University of Alberta

POLYSEARCH: A WEB BASED TEXT MINING SYSTEM FOR EXTRACTING RELATIONSHIPS BETWEEN HUMAN DISEASES, GENES, MUTATIONS, DRUGS AND METABOLITES

by

Dean Hsu Chou Cheng ©

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of **Master of Science**.

Department of Computing Science

Edmonton, Alberta
Fall 2007

Library and
Archives Canada

Published Heritage
Branch

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque et
Archives Canada

Direction du
Patrimoine de l'édition

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence
ISBN: 978-0-494-33217-7
Our file Notre référence
ISBN: 978-0-494-33217-7

NOTICE:
The author has granted a non-exclusive license allowing Library and Archives Canada to reproduce, publish, archive, preserve, conserve, communicate to the public by telecommunication or on the Internet, loan, distribute and sell theses worldwide, for commercial or non-commercial purposes, in microform, paper, electronic and/or any other formats.

The author retains copyright ownership and moral rights in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

AVIS:
L'auteur a accordé une licence non exclusive permettant à la Bibliothèque et Archives Canada de reproduire, publier, archiver, sauvegarder, conserver, transmettre au public par télécommunication ou par l'Internet, prêter, distribuer et vendre des thèses partout dans le monde, à des fins commerciales ou autres, sur support microforme, papier, électronique et/ou autres formats.

L'auteur conserve la propriété du droit d'auteur et des droits moraux qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

In compliance with the Canadian Privacy Act some supporting forms may have been removed from this thesis.

While these forms may be included in the document page count, their removal does not represent any loss of content from the thesis.

Conformément à la loi canadienne sur la protection de la vie privée, quelques formulaires secondaires ont été enlevés de cette thèse.

Bien que ces formulaires aient inclus dans la pagination, il n'y aura aucun contenu manquant.

Canada

Abstract

Scientists are deluged with information with more than 1,000,000 scientific articles being published a year. Almost half of them are biomedical in nature. Not only is there an overwhelming abundance of scientific articles, there is an overwhelming abundance of biomedical information in electronic database such as OMIM, SwissProt, DrugBank and others.

One way to address this "information overload" is to develop computational tools to extract relevant information. In this thesis, we present a web-based biomedical text mining system named PolySearch, that has been designed to extract relationships between human diseases, genes, mutations, drugs, and metabolites, from a variety of text sources and electronic databases. PolySearch allows diverse search and text ranking possibilities not found in most other biomedical text mining tools. We also demonstrate that PolySearch is able to achieve a high level of performance in comparison to other biomedical text mining tools. The server is freely available at http://wishart.biology.ualberta.ca/polysearch.

Acknowledgements

I would like to thank everyone that I met in Edmonton for providing all the quality time that I had here. Special thanks to my supervisor Dr. David Wishart for providing valuable supervision for my research and my committee members, Dr. Russ Greiner and Dr. Sambasivarao Damaraju, for helpful suggestions to my thesis. Thanks to Craig Knox, Haiyan Zhang, Nelson Young, Summit Sawhney, and Dr. Paul Stothard for helping me on this project. I would also like to thank my friend Tom for asking me to come to the University of Alberta to study, and my friends Ben and Xian for being "fellow graduate students". Finally I want to thank all my family members for providing their support during my studies.

Table of Contents

1. Introduction........1
2. Background and Related Work........7
  2.1 Background........7
  2.2 Nature Language Processing........10
  2.3 Related Work........17
3. PolySearch System Overview........35
  3.1 Overview........35
  3.2 Query Interface........37
  3.3 PolySearch's Databases........42
  3.4 Custom Thesauruses........46
4. PolySearch Data Mining........49
  4.1 PolySearch's Text Mining System........49
  4.2 PolySearch Sentence Scoring, Ranking and Integration........53
  4.3 PolySearch Results........60
  4.4 Improve Association Word Selections/Relevance Feedback........66
  4.5 SNPs........71
  4.6 Primer Design........72
5. Evaluation........75
  5.1 Results........75
    5.1.1 Feature Comparison........76
    5.1.2 Evaluation of Gene/Protein Synonym Identification........79
    5.1.3 Evaluation of Protein-Protein Interactions........80
    5.1.4 Evaluation of Drug/Gene Associations........89
    5.1.5 Evaluation of Metabolite/Gene Associations........95
    5.1.6 Evaluation of Disease/Gene Associations........98
    5.1.7 Manual versus Automated........101
    5.1.8 Final Example........103
  5.2 Discussion........107
6. Conclusion and Future Work........113
  6.1 Conclusion........113
  6.2 Future Work........114
Bibliography........117

List of Tables

Table 3.1: A detailed listing of all allowed "basic" queries in PolySearch. 38
Table 3.2: Statistics for PolySearch's thesauruses. 48
Table 5.1: Feature comparison of various biomedical text mining tools. 78
Table 5.2: Precision, recall and f-measure on gene synonym identification for PolySearch and IHOP. 79
Table 5.3: Precision, recall and f-measure on a corpus of protein-protein interaction using PolySearch's rule based pattern recognition system (when X is given) and using a Naïve Bayes classifier available from Weka. 82
Table 5.4: The five SwissProt IDs, gene symbols, and gene names randomly chosen as input to evaluate how different tools perform for protein-protein interaction. 85
Table 5.5: Precision, recall, and f-measure for protein-protein interaction evaluation among the different tools. 86
Table 5.6: The DrugBank IDs and common names for the ten drugs randomly chosen from DrugBank for evaluating "Given Drug Find Associated Gene" queries. 90
Table 5.7: "Given Drug Find Associated Gene": comparing DrugBank, LitMiner, EBIMed, PolySearch with PubMed, and PolySearch with PubMed + DrugBank. 90
Table 5.8: Precision for drug-gene associations of the ten "Given Drug Find Associated Gene" queries at different R1 scores and precision for all R1 sentences of the ten queries. 93
Table 5.9: The HMDB IDs and common names for the ten metabolites randomly chosen from HMDB for evaluating Given Metabolite Find Associated Gene queries. 96
Table 5.10: "Given Metabolite Find Associated Gene": precision, recall and f-measure for HMDB, LitMiner, EBIMed, PolySearch with PubMed, PolySearch with PubMed + OMIM, and PolySearch with PubMed + OMIM + HMDB. 97
Table 5.11: Precision for metabolite-gene associations of the ten "Given Metabolite Find Associated Gene" queries at different R1 scores and precision for all R1 sentences of the ten queries. 98
Table 5.12: The disease names for the ten diseases randomly chosen for evaluating Given Disease Find Associated Gene queries. 99
Table 5.13: "Given Disease Find Associated Gene": precision, recall and f-measure for GAD, LitMiner, EBIMed, PolySearch R2 >= 1, PolySearch R1 >= 1, PolySearch with PubMed + OMIM, and PolySearch with PubMed + OMIM + GAD. 100
Table 5.14: Precision for disease-gene associations of the ten "Given Disease Find Associated Gene" queries at different R1 scores and precision for all R1 sentences of the ten queries. 101

List of Figures

Figure 2.1: The results of using COX2 (and its synonyms) and Paclitaxel (and its synonyms) as the query for MedMiner's Gene-Drug search. 18
Figure 2.2: The list of genes returned by MedGene for the disease query "Colonic Neoplasms" and choosing product of frequency as the statistical method. 20
Figure 2.3: LitMiner's Disease – Gene output for colon cancer. 23
Figure 2.4: Output from ALIBABA using colon cancer as the query and maximum number of abstracts is set as 100. 24
Figure 2.5: Output from IHOP using COX-2 as the query. 27
Figure 2.6: An IHOP abstract used to find COX-2 gene interactions. 28
Figure 2.7: Output from EBIMed using "colon cancer" as the query, searching for human genes and maximum number of abstracts is set to 2000. 32
Figure 2.8: Key sentences found by EBIMed for associations between colon cancer and COX-2. 33
Figure 3.1: PolySearch system overview showing the resources that PolySearch uses and the features found in PolySearch. 36
Figure 3.2: PolySearch's homepage where users can select the different "given X find associated Y's" queries. 37
Figure 3.3: The query refinement page for "Given Disease Find Associated Gene". 39
Figure 4.1: An example of each of the R sentence. The query is colon cancer, the filter words are coloured in fuchsia, and p53 is the protein of interest. 58
Figure 4.2: An example of the output for PolySearch's main results display. 60
Figure 4.3: An example of the key sentences that are extracted and evaluated from a standard PolySearch run. 63
Figure 4.4: The key sentences EBIMed found for the query N-acetyl-D-glucosamine and gene of interest is NDST-1. 65
Figure 4.5: The key sentences PolySearch found for the query N-acetyl-D-glucosamine and gene of interest is NDST-1. 66
Figure 4.6: Results from a "Given Text Word Find Metabolites" query where the query is cerebrospinal fluid, the filter words are metabolite, metabolites, compound and compounds, and the maximum number of abstracts is set to 2000. 68
Figure 4.7: The key sentences for lactate for the "Given Cerebrospinal Fluid Find Metabolites" search. 69

Figure 4.8: Some example R1 or R2 sentences that PolySearch found for the "Given Cerebrospinal Fluid Find Metabolites" search. These examples were found while briefly browsing through the results. 70
Figure 4.9: The output for a "SNP to Gene" search using rs2234953, rs2266633, rs2266636, and rs2266637 as input. PolySearch collects important information about SNP such as: position, type of polymorphism, gene symbol/name, function and allelic frequency. 72
Figure 4.10: An illustration of the PCR Primer Design feature in PolySearch. 74
Figure 5.1: Results from a "Disease to Gene" query where the query is colon cancer, the filter words are SNP, SNPs, Polymorphism and Polymorphisms, the minimum R2 filter is set to 1, and the maximum number of abstracts is set to 2000. 104
Figure 5.2: The key sentences for SNP/polymorphism association between colon cancer and GSTT1. 105
Figure 5.3: The output for a "Gene to SNP" search using GSTT1 as input. PolySearch collects important information about SNP such as: position, type of polymorphism, gene symbol/name, function and allelic frequency 106
Figure 5.4: An illustration of the PCR Primer Design feature in PolySearch. 107

Chapter 1

1. Introduction

Today's scientists are deluged with information. Currently there are more than 8000 scientific, technical and medical journals publishing more than 1,000,000 articles a year. Nearly 50% of these articles are biomedical in nature. Indeed, it has been estimated that in order for a scientist to stay current for a single high-priority disease (say breast cancer), they would have to scan 130 different journals and read 27 papers each week [1]. Given that most journal articles are not exactly "light" reading, this task of staying current with the literature could easily occupy 75% of a scientist's working day.

The problem with information overload is not restricted to scientific papers. Electronic databases are equally culpable. Thousands of web-accessible text, image and sequence databases now exist [2]. These contain terabytes of data and are expanding in both number and size far faster than the rate of scientific publishing. For instance, GenBank, which doubles in size every 12 months, contains 60 million sequences occupying 250 Gigabytes (www.ncbi.nlm.nih.gov/Genbank/index.html). Just tracking the appearance and content of new databases, let alone using the information in them, can prove to be a full time challenge. Equally problematic is the task of checking that the data in the sequence, structure, drug and gene expression databases is current with the information in the literature (and vice versa).

Clearly, the quantity of information generated by the scientific community is far too great for any human to efficiently process or assimilate. Too much fragmentary information and non-contextual data exists in too many places. This makes the task of finding relevant information on a specialized topic somewhat like finding a proverbial needle in the haystack. It is now obvious that a key challenge, especially in the field of bioinformatics, is to develop methods that allow this information to be easily found and readily exploited by human users.

An important advance in this area has come with the development of NCBI's new Entrez Cross-Database search system [3]. This information retrieval system brings the hunt for new and useful biomedical data to a new level by integrating PubMed (i.e. biomedical abstract data) with NCBI's multitude of sequence, structure and chemical databases. Now users can enter a term, such as "breast cancer" and almost instantly see a hyperlinked list of journal abstracts (162,000) containing the term, genes and proteins associated with breast cancer (courtesy of GenBank), 3D structures of proteins associated with the disease (from Entrez Structure and MMDB), drugs or drug candidates used to treat breast cancer (from PubChem) along with microarray data (from GEO), SNP information (from Entrez SNP) as well as links to another 10 NCBI database resources (GENSAT, STS, UniGene, OMIM, etc.). Entrez is a superb resource that greatly improves the speed and precision with which researchers can find relevant data on a given gene, disease, mutation, drug or microarray experiment.

However, Entrez is still somewhat limited because it is restricted to searching its abstract and molecular database resources through MeSH (Medical Subject Heading) terms, MeSA (Medical Subject Annotation) terms and keywords in database titles or database names. In other words, Entrez doesn't look through the text of all 162,000 abstracts on breast cancer, nor can it find a list of genes that are mentioned in those abstracts, extract key sentences for those genes, count the frequency of appearance of those genes and provide a frequency or relevancy ranking for them. Likewise, Entrez does not link its results to many equally useful external databases such as SwissProt [4], Human Gene Mutation Database (HGMD) [5], DrugBank [6] or the Human Metabolome Database (HMDB) [7]. Another unfortunate limitation is that Entrez does not contain disease, gene/protein, drug or metabolite thesauruses. For instance, if one wanted to find all the drugs that could be used to treat cancer, one would have to repeatedly enter "breast cancer AND Y" where Y is the name of each of the 25,000 known drugs, brand names and their synonyms. In other words, Entrez does not have a pre-assembled list of all 25,000 known drug names/synonyms and it does not search for co-occurrences of those drugs with the words "breast cancer" in PubMed abstracts.

These kinds of sophisticated text searching tasks are more suited to a different class of programs called medical text mining systems. Several excellent biomedical text mining tools now exist such as MedMiner [8], MedGene [9], LitMiner [10], iHOP [11], ALIBABA [12] and EBIMed [13]. These tools exploit the explicit textual information contained within the PubMed database by selecting or highlighting key sentences or terms within the abstracts and then summarizing or presenting the results in some form. The text mining capabilities of MedMiner, MedGene, LitMiner, iHOP, ALIBABA and EBIMed greatly exceed what one can do or view with Entrez/PubMed. However, these text mining tools were designed specifically to extract information only from PubMed abstracts and no other databases (i.e. OMIM, DrugBank). Ideally what is needed is something that combines the text mining capabilities found in MedMinder, MedGene, LitMiner, iHOP, ALIBABA and EBIMed with some of the database integration found in Entrez. What's more, one would like to see some analytical capabilities built into such a system so that users could manipulate, view, or archive the resulting information (text or sequence) in a convenient, web-accessible format. These requirements motivated us to develop just such a resource – called PolySearch.

PolySearch is a web-accessible tool that is designed specifically for extracting and analyzing the relationship between human diseases, genes/proteins, mutations (SNPs), drugs, metabolites, tissues, organs, and subcellular localizations. It extracts and analyzes not only PubMed data, but also data from multiple databases (OMIM, DrugBank, SwissProt, HGMD, Entrez SNP, etc.) using sophisticated data mining tools. It also displays, links and ranks text, as well as sequence data in multiple forms and formats. PolySearch differs from other tools in the number of search possibilities, its general search strategies, its support for data analysis (relevance ranking based on frequency, co-occurrence, association words and pattern recognition, SNP analysis and primer design), its large collection of customized thesauruses and its extensive integration with other databases. In other words PolySearch tries to combine the best of Entrez with the best of MedMiner, MedGene, LitMiner, iHOP, ALIBABA and EBIMed. By comparing PolySearch to other biomedical text mining tools, we will also demonstrate that it is one of the more effective tools in extracting relevant sentences to facilitate extraction of meaningful biological associations. PolySearch allows users to read a set of organized relevant sentences for each association rather than reading the abstracts themselves. In particular, users also have direct control over what PolySearch considers as relevant. Through the use of a unique scoring scheme, "the PolySearch Relevancy Index", PolySearch provides visual cues to facilitate rapid assimilation of association strength, as well as a means for automatic information extraction. Overall PolySearch is a novel biomedical text mining system that, with its diverse search and text ranking possibilities, provides new and more effective biomedical text mining capabilities.

This thesis describes the design, implementation and testing of PolySearch. Specifically, chapter 2 provides a brief background on biomedical text mining and describes some of the challenges that are faced in the field of biomedical text mining. This chapter also gives an overview on related biomedical text mining tools available for extracting information from biomedical text abstracts. In chapters 3 and 4, we present details on how the system was designed and its proposed methods for extracting information. Chapter 5 describes a range of assessments that were carried out to evaluate and demonstrate the capabilities of PolySearch's different search possibilities. In Chapter 6 the thesis ends with a brief conclusion and a description of possible future work. The primary approach taken in constructing the PolySearch system is a manually crafted rule-based methodology. In the future, PolySearch can benefit from incorporating systematic artificial intelligence or machine learning approaches.

Thesis statement: We demonstrate that it is possible to use a sentence-based approach in combination with manually curated thesauruses and a 4-state information content measure for informative sentences, to improve the quality and scope of biomedical information retrieval. It is also hypothesized that by integrating high quality, curated biomedical databases with journal abstract information that it is possible to improve the coverage and precision of data retrieval over what is currently achieved by other methods or tools.

Chapter 2

2. Background and Related Work

2.1 Background

The language of biology, like natural languages, is often unstructured. However biological or biomedical text can be particularly difficult to understand as many specialized terms (i.e. jargon) are commonly used throughout the existing literature. The following is an example of an abstract [14] in PubMed about colon cancer:

> Loss of AP-2alpha results in deregulation of E-cadherin and MMP-9 and an increase in tumorigenicity of colon cancer cells in vivo. Activator protein-2 (AP-2) is a transcription factor that regulates proliferation and differentiation in mammalian cells and has been implicated in the acquisition of the metastatic phenotype in several types of cancer. Herein, we examine the role of AP-2alpha in colon cancer progression. We provide evidence for the lack of AP-2alpha expression in the late stages of colon cancer cells. Re-expression of the AP-2alpha gene in the AP-2alpha-negative SW480 colon cancer cells suppressed their tumorigenicity following orthotopic injection into the cecal wall of nude mice. The inhibition of tumor growth could be attributed to the increased expression of E-cadherin and decreased expression and activity of matrix-metalloproteinase-9 (MMP-9) in the transfected cells, as well as a substantial loss of their in vitro invasive properties. Conversely, targeting constitutive expression of AP-2alpha in AP-2-positive KM12C colon cancer cells with small interfering RNA resulted in an increase in their invasive potential, downregulation of E-cadherin and increased expression of MMP-9. In SW480 cells, re-expression of AP-2alpha resulted in a fourfold increase in the activity of E-cadherin promoter, and a 5-14-fold decrease in the activity of MMP-9 promoter, indicating transcriptional regulation of these genes by AP-2alpha. Chromatin immunoprecipitation assay showed that re-expressed AP-2alpha directly binds to the promoter of E-cadherin, where it has been previously reported to act as a transcriptional activator. Furthermore, chromatin immunoprecipitation assay revealed AP-2alpha binding to the MMP-9 promoter, which ensued by decreased binding of transcription factor Sp-1 and changes in the recruitment of transcription factors to a distal AP-1 element, thus, contributing to the overall downregulation of MMP-9 promoter activity. Collectively, our data provide evidence that AP-2alpha acts as a tumor suppressor gene in colon cancer.

To capture relevant information from this abstract, one first has to have a sufficient background in biology to know that the terms “AP-2alpha”, “E-cadherin” and “MMP-9” are names of genes or proteins. Further, one needs a good knowledge of molecular biology to understand that “AP-2alpha” is a transcription factor that regulates the expression of “E-cadherin” and “MMP-9” thereby playing a role in suppressing colon cancer. Upon reading through the abstract, the reader may realize that “AP-2alpha”, “Activator protein-2”, and “AP-2” are in fact the same gene/protein, meaning that three different synonyms were used in this abstract alone. This demonstrates one of the challenges in understanding biomedical literature as many of the specialized terms also have a large number of synonyms. For instance a gene and the proteins coded by this gene often have different names, yet they are synonymous to each other. In addition, a given gene may have multiple synonyms while the protein coded by this gene may have a different set of synonyms that are, in turn, all synonyms of each other. As we have seen in this example, three different names of the protein AP-2alpha have been used. In fact TFAP2A, which is the official gene symbol for AP-2alpha recommended by the Human Genome Organisation Gene Nomenclature Committee (HGNC [15]), is nowhere to be found in the abstract. So if one restricts the search to finding the official gene symbol, TFAP2A, one would miss the important associations between AP2-alpha and E-cadherin and MMP-9. Alternately, if another abstract only uses TFAP2A and it further shows evidence of association between TFAP2A and MMP-9, without data indicating that TFAP2A, AP-2alpha, Activator protein-2 and AP-2 are synonyms then instead of two abstracts showing associations between TFAP2A and MMP-9, only one abstract would appear to state this relationship.

Another challenge in understanding biomedical text lies in the inherent complexity of biological systems. Living systems are composed of tens of thousands of genes, tens of thousands of proteins, thousands of metabolites, hundreds of different types of cells/tissues and dozens of different cell types. All of these components interact in innumerable ways that cannot easily be described in one or two-word phrases. Similarly the diseases and disorders that arise when these components are broken or missing also cannot be described with simple words or descriptions. Indeed, physicians, pharmacists, and biologists spend a significant portion of their formal education on learning or memorizing the terminology and phrases used to describe what they are seeing, measuring or diagnosing.

The primary goal of biomedical text mining is to provide computational tools that make the understanding of biomedical texts easier. More specifically, text mining must: 1) support human experts in extracting relevant information and 2) facilitate automatic information extraction such as "given X find all associated Ys". Building biomedical text mining tools cannot be done in a vacuum. Indeed, inputs from biological experts are essential in making a proper tool since computing scientists generally lack the knowledge needed to understand the complex information we find in biomedical texts. In cases where the text mining tool is drawing conclusions or making assertions, it is important to properly convey how and why the tool is asserting that X is associated with Y.

Properly conveying how and why serves three purposes. First of all, exactly how or why X is associated with Y is what biologists really want to know in the first place, not just the fact that X is associated with Y. Second, the provision of explanations garners the trust of biologists in using the tool by supporting assertions with the reasoning or evidence behind them. Third, the presentation of both facts and explanations provides a means for biologists to give feedback to the computing scientists to make the text mining tool better and more relevant. These are the key issues to keep in mind when developing and evaluating a biomedical text mining tool. Many of today's better text mining tools take these three issues into account in some way or another.

2.2 Nature Language Processing

Natural Language Processing (NLP) is an area of research that combines natural language linguistics and computing science in an effort to ultimately allow computers to understand languages as humans do. There are many applications in NLP including speech recognition, machine translation, question-answering, semantic web applications, document summarization, and information retrieval. A large number of NLP applications share the same challenges in evaluating the performance of different systems and in developing methods to achieve their goals. As a result, the lessons learned in NLP can potentially benefit many research efforts in biomedical text mining. For example, "name entity recognition" is a term used in NLP to denote finding person names, location names, organization names, and other proper names in newswire articles. The techniques developed in name entity recognition could potentially be applied to finding disease/gene/drug/metabolite names in biomedical texts. This section takes a closer look at a field of NLP that most applies to biomedical text mining, namely, information retrieval.

Information Retrieval

A simple information retrieval system is a system that retrieves a ranked set of relevant documents related to a given query from a document collection. For example, PubMed is an information retrieval system such that when given a query like "Colon Cancer", it will return a list of abstracts mentioning "Colon Cancer" ranked by the date of publication. As the amount of information continues to grow, retrieving only relevant documents becomes insufficient as users want and expect more of a text mining system. For instance, many of today's users want their text mining systems to process document text and retrieve relevant passages (phrases, sentences, paragraphs, etc.) thereby directly "answering" the query. The distinction between information retrieval systems and text mining systems continues to blur as modern information retrieval systems often contain text mining components. In addition, text mining systems can often be used as information retrieval systems since documents that contain relevant passages should be more relevant. Therefore, the evaluation methodologies developed for information retrieval systems can easily be applied to evaluate the performance of text mining systems. To this end, there is a well known initiative, TREC (Text REtrieval Conference), which is designed to foster research on technologies for information retrieval. TREC has four main goals [16]:

- to encourage retrieval research based on large test collections;
- to increase communication among industry, academia, and government by creating an open forum for the exchange of research ideas;
- to speed the transfer of technology from research labs into commercial products by demonstrating substantial improvements in retrieval methodologies on real-world problems; and
- to increase the availability of appropriate evaluation techniques for use by industry and academia including development of new evaluation techniques more applicable to current systems.

TREC provides large test collections for several different information retrieval tasks in different fields and allows groups and individuals from all over the world to participate in the tasks. Each test collection contains three components: 1) documents, 2) topics for the documents, and 3) relevancy judgments where each document is indicated to be either relevant to the document's topic or irrelevant to the document's topic. The relevancy judgments were completed by a team of assessors who are experts in the field in which the documents belong. When a collection of documents is too large for complete manual relevancy judgments of all the documents in the collection, pooling is done first and then all the pooled documents are judged to be relevant or irrelevant. Pooling is a technique where the top X documents from each participant's results are pooled together forming a subset of the original collection that is easier to manage [17].

The use of pooling to produce a test collection has been questioned and it has been noted that for pooling to be valid, there needs to be enough relevant documents in the pool and the relevant documents must be unbiased [16]. Once a proper test collection is constructed with relevancy judgment, several performance measures are used to evaluate a system's performance including precision, recall and f-measure. Precision is a measure of the accuracy of the system, i.e. how many of the system's extracted results are true.

$$precision = \frac{true\ positive}{true\ positive + false\ positive}$$

Recall is a measure of the coverage of the system (i.e. did the system miss any meaningful results?)

$$recall = \frac{true\ positive}{true\ positive + false\ negative}$$

F-measure (*f*) is a combination score of precision and recall.

$$f = \frac{2 \times precision \times recall}{precision + recall}$$

The maximum of all three measures is 1 (i.e. 100%), which is the ideal situation. However, there is typically a trade off between precision and recall. One can achieve 100% recall by predicting that, for instance, one gene interacts with all the genes in the human body. However, such a prediction would have very low precision. One can also try to be very precise and then miss many possible true associations. Therefore, f-measure is used as a composite score between precision and recall in order to achieve fair comparison of systems with different precisions and recalls.

Average precision is another single-valued measure that has both precision and recall components. Average precision (AP) favors systems that rank relevant results higher and it can be calculated using the following formula:

$$AP = \frac{\sum_{R=1}^{N} \Pr(R) * rel(R)}{true\ positive}$$

where R is the rank, N is the number of results retrieved, Pr(R) is the precision of the Rth result and rel(R) is 1 if the Rth result is true, 0 if the Rth result is false. For example, if there are a total of 3 true positives and a system retrieves them as the first 3 results, then AP is

$$\frac{(1/1 + 2/2 + 3/3)}{3} = 1$$

However, if a system's third result is false and the fourth result is true, then AP is

$$\frac{(1/1 + 2/2 + 3/4)}{3} = 0.917$$

When there are multiple queries, with systems returning a ranked list of results for each query, then mean average precision (the mean of the average precisions of different queries) is often used as the single-valued performance measure to compare different systems.

Relevance Feedback/Query Refinement

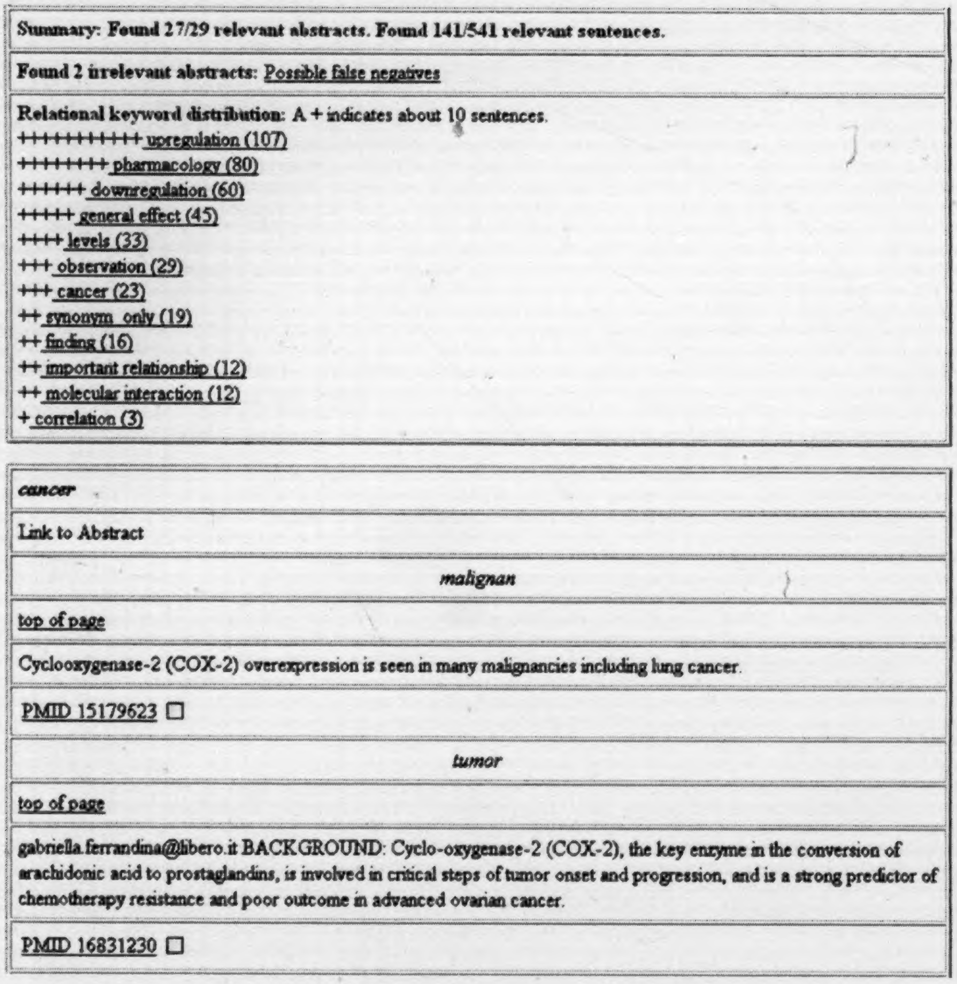
While the previous section described how one evaluates information retrieval systems when relevancy judgments have already been done for each document in the test collection, one detail not mentioned is how to determine whether a document is relevant or not. When building an information retrieval/text mining system, one has to anticipate what is relevant to the users and then build a system that can distinguish between what is relevant and what is irrelevant in order to be effective. However, anticipating what is relevant to the users is often hard to do as different users are interested in different things and sometimes even the users do not know what is relevant to them until they see the available results returned by a system. This has led to the development of relevance feedback techniques in information retrieval/text mining systems. Relevance feedback occurs after the system displays its initial results, particularly when a user is asked to identify the relevant documents/keywords in the initial result set. These relevant documents/keywords identified by the user are then used to refine the query (or refine the system's relevancy ranking) leading to a second result set where relevance results should be ranked higher compared to the initial result set. The general strategy for query refinement through relevance feedback is to add more relevant keywords to the query such that the new expanded query can be matched better to relevant documents through some similarity measures. There are many similarity measure models that can be used. For example, a pure occurrence model is one where the occurrence of a query term in a document makes the document match to the query. A vector space model is one where the query term and document are represented as vectors of "term-frequency weighting scores" and then the vectors can be used to calculate the similarity between the query and a document.

In general, relevance feedback techniques can be divided into two main approaches: automatic or manually interactive. For automatic query refinement, the relevance feedback system automatically generates the list of relevant keywords to be appended to the query for the next iteration of relevance feedback. On the other hand, for manual interactive query refinement, the relevance feedback system asks users to provide relevant keywords to expand the query. There are advantages and disadvantages to both approaches. It has been reported that automatic query refinement can improve overall performance of an information retrieval system; however, the magnitude of improvement varies from different queries and different document collections [18]. Manual interactive query refinement from an expert can achieve even greater (and more consistent) improvement compared to automatic query refinement. However, the performance improvement generated from manual interactive query refinement of a novice relevance feedback user is generally inferior to the improvement provided by automatic query refinement [18]. Overall, relevance feedback can improve the performance of an information retrieval/text mining system. Therefore providing some form of relevance feedback mechanism in an information retrieval/text mining system would be very useful.

The main objective for a text mining system is to identify or extract relevant information in order to save time that would otherwise be spent reading abstracts one at a time. Different tools provide different ways of achieving this goal. The next section provides a more detailed look into some of the better known biomedical text mining tools.

2.3 Related Work

MedMiner was one of the earliest web-based biomedical text mining tools to be developed [8]. It provides support for “Gene only” searches, “Gene to Gene” searches, “Gene to Drug” searches and “General Query” searches. The “Gene to Gene” search and the “Gene to Drug” search allow only single component queries (one gene to one gene search and one gene to one drug search respectively), as opposed to one to many kinds of queries. For a given query, MedMiner allows the users to send the query first to GeneCards [19] returning a list of genes that users can choose before proceeding to querying PubMed. This GeneCards filter step can be bypassed for non-gene queries. In addition, once the query gene or query drug is chosen, MedMiner allows users to select gene synonyms or drug synonyms to make a combined gene synonym or drug synonym query to PubMed. Once the abstracts are retrieved from PubMed, MedMiner tries to highlight the relevant sentences within the abstracts that contain the name of the query gene or the query drug plus a keyword from a pre-defined list of keywords such as: inhibit, block, report, tumor, result, etc. The results are organized into twelve general categories followed by the relevant sentences (Figure 2.1).



Summary: Found 27/29 relevant abstracts. Found 141/541 relevant sentences.

Found 2 irrelevant abstracts: Possible false negatives

Relational keyword distribution: A + indicates about 10 sentences.
+++++++++++ upregulation (107)
++++++++ pharmacology (80)
++++++ downregulation (60)
+++++ general effect (45)
++++ levels (33)
+++ observation (29)
+++ cancer (23)
++ synonym_only (19)
++ finding (16)
++ important relationship (12)
++ molecular interaction (12)
+ correlation (3)

cancer

Link to Abstract

malignan

top of page

Cyclooxygenase-2 (COX-2) overexpression is seen in many malignancies including lung cancer.

PMID 15179623 ☐

tumor

top of page

gabriella.ferrandina@libero.it BACKGROUND: Cyclo-oxygenase-2 (COX-2), the key enzyme in the conversion of arachidonic acid to prostaglandins, is involved in critical steps of tumor onset and progression, and is a strong predictor of chemotherapy resistance and poor outcome in advanced ovarian cancer.

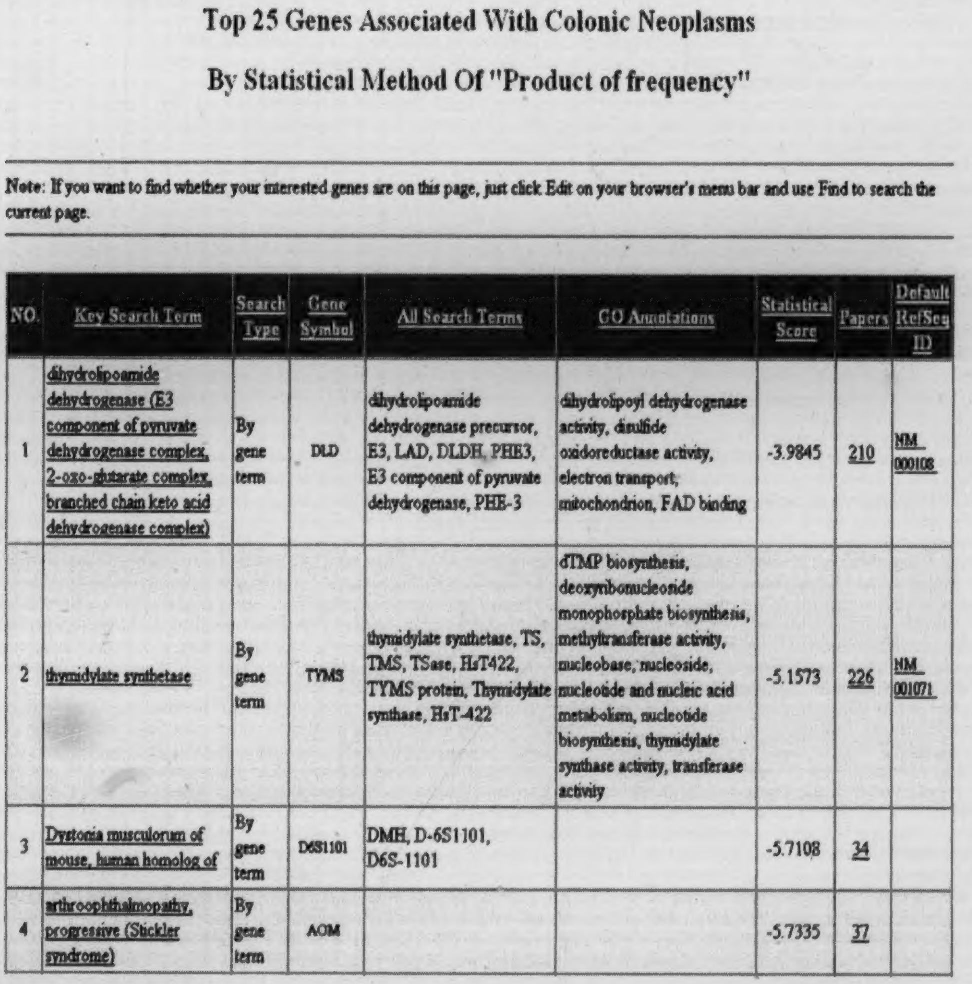
PMID 16831230 ☐

Figure 2.1: The results of using COX2 (and its synonyms) and Paclitaxel (and its synonyms) as the query for MedMiner's Gene-Drug search.

Evaluating a biomedical text mining tool can be difficult as it is hard to measure the most important question: "how much time does using a biomedical text mining tool save?" However, through thorough analysis of several queries, the authors of MedMiner did show that this relevant sentence highlighting approach properly identifies relevant abstracts without omitting abstracts that have relevant information. Furthermore, the investigators within MedMiner's group have largely opted to use MedMiner instead of PubMed as they felt that using MedMiner offers a greater time saving and that it is easier to digest the results by reading structured set of sentences rather than full abstracts [8].

In MedMiner, because a relevant sentence is defined as a sentence that contains at least one gene or one drug plus a keyword, the relevant sentences may not show any association between the gene and the drug since the relevant sentences may mention only the gene or the drug. There is no preference or added score for a relevant sentence mentioning both the gene and the drug in the same sentence. As a result, to find relevant sentences containing both terms, users must perform a manual search. The major limitation of MedMiner is that it only supports single component queries making it a somewhat more specialized tool for biologists knowing what they want to study and less of an automatic information extraction tool.

MedGene provides disease-gene (given a disease, find all associated genes) and gene-gene (given a gene, find all associated genes) searches [9]. MedGene works on finding word co-occurrences at the abstract level. For each disease-gene pair, MedGene first defines a contingency table of four numbers including: 1) the number of abstracts with both disease X and gene Y (disease/gene double hits), 2) the number of abstracts with disease X (disease single hits), 3) the number of abstracts with gene Y (gene single hits) and 4) the number of abstracts with neither disease X nor gene Y. Using the four numbers in this contingency table, MedGene provides several statistical options (product of frequency, probability, chi square analysis, Fischer exact test, relative risk of gene, relative risk of disease) to estimate the strength of the association between disease X and gene Y. The term "product of frequency" for example, is the product of the proportion of disease/gene double hits to disease single hits and the proportion of disease/gene double hits to gene single hits. An example output from MedGene is shown in Figure 2.2.



Top 25 Genes Associated With Colonic Neoplasms

By Statistical Method Of "Product of frequency"

Note: If you want to find whether your interested genes are on this page, just click Edit on your browser's menu bar and use Find to search the current page.

NO.	Key Search Term	Search Type	Gene Symbol	All Search Terms	GO Annotations	Statistical Score	Papers	Default RefSeq ID
1	dihydrolipoamide dehydrogenase (E3 component of pyruvate dehydrogenase complex, 2-oxo-glutarate complex, branched chain keto acid dehydrogenase complex)	By gene term	DLD	dihydrolipoamide dehydrogenase precursor, E3, LAD, DLDH, PHE3, E3 component of pyruvate dehydrogenase, PHE-3	dihydrolipoyl dehydrogenase activity, disulfide oxidoreductase activity, electron transport, mitochondrion, FAD binding	-3.9845	210	NM_000108
2	thymidylate synthetase	By gene term	TYMS	thymidylate synthetase, TS, TMS, TSase, HsT422, TYMS protein, Thymidylate synthase, HsT-422	dTMP biosynthesis, deoxyribonucleoside monophosphate biosynthesis, methyltransferase activity, nucleobase, nucleoside, nucleotide and nucleic acid metabolism, nucleotide biosynthesis, thymidylate synthase activity, transferase activity	-5.1573	226	NM_001071
3	Dystonia musculorum of mouse, human homolog of	By gene term	D6S1101	DMH, D-6S1101, D6S-1101		-5.7108	34	
4	arthroophthalmopathy, progressive (Stickler syndrome)	By gene term	AOM			-5.7335	37	

Figure 2.2: The list of genes returned by MedGene for the disease query "Colonic Neoplasms" and choosing product of frequency as the statistical method.

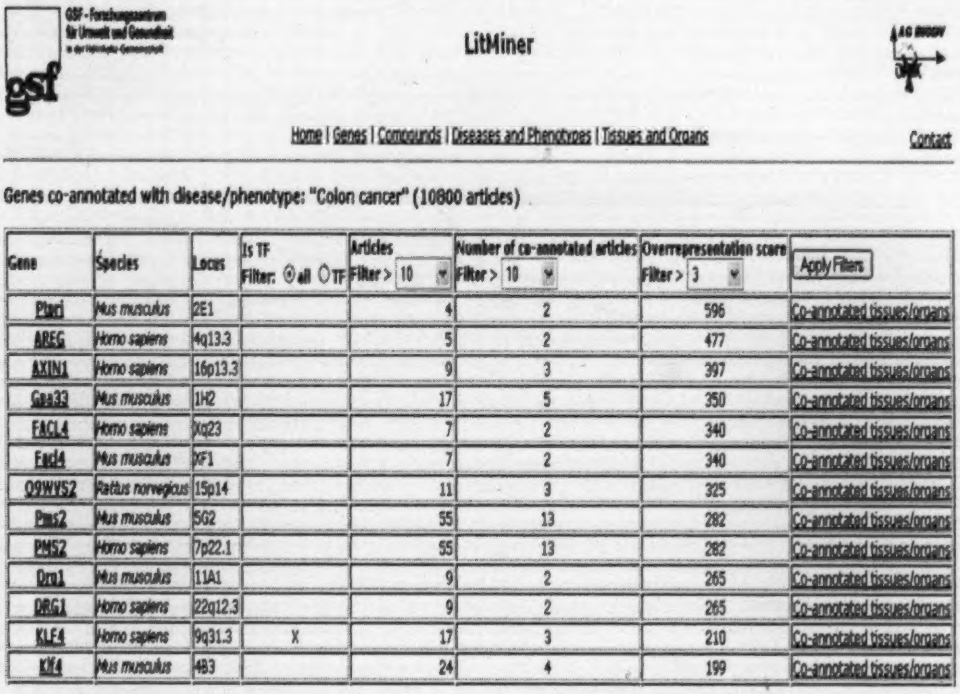
In evaluating MedGene's performance, the authors took a closer look at a single disease, prostate cancer. For this particular condition, they electronically retrieved and manually analyzed the abstracts for the highest ranked 100 genes and the lowest ranked 200 genes related to prostate cancer from MedGene. Their analysis showed that 77.5% of the highest ranked 100 genes fell into one of the five categories reflecting meaningful gene-disease relationships [20]. In the same analysis, the authors also reported that 67.4% of the lowest ranked 200 prostate cancer genes from MedGene reflected true relationships. In another analysis, MedGene's output for a breast cancer query was compared with microarray gene expression data for breast cancer and normal breast tissue samples. The analysis suggested that genes with a more substantial expression level in microarray data were more likely to have a stronger breast cancer association in the literature. The textual analysis of disease-gene associations could potentially complement microarray gene expression analysis by suggesting which genes are more important. It may also allow researchers to better understand why gene expression changes occur under different conditions.

While MedGene does provide statistical rankings for disease/gene associations and while it does provide hyperlinks to PubMed abstracts of interest, MedGene offers no text or sentence highlighting capabilities. As can be seen in Figure 2.2, the hyperlinks under the column heading "Papers" (the eighth column) link back to PubMed where it lists unprocessed abstracts. In other words, unlike MedMiner, MedGene does not highlight key words or select informative sentences. So, for example, if one wished to understand how and why colonic neoplasms are associated with the DLD gene, a user would have to manually select and read through all 210 abstracts themselves to obtain the information they wanted. The only evidence that MedGene provides for its associations is a statistical score. This statistical score may not represent a true biological relationship and to use MedGene as an automatic information extraction tool requires a clear understanding of what is statistically significant and what isn't. Unfortunately there is no clear or recommended threshold (as is often used in BLAST scores), so it would be difficult to use MedGene as an automatic information extraction tool.

LitMiner is another example of a biomedical text mining system that uses the word co-occurrence approach [10]. LitMiner predicts relationships between genes, chemical compounds, diseases and tissues. LitMiner calculates what is called an over-representation score based on: 1) the total number of abstracts examined (TNA); 2) the number of abstracts with both disease X and gene Y (NCO(KT1-KT2); 3) the number of abstracts with disease X (NA(KT1)), and 4) the number of abstracts with gene Y (NA(KT2)). The over-representation score is calculated using the following equation.

$$over-representation = \frac{TNA \times NCO(KT1 - KT2)}{NA(KT1) \times NA(KT2)}$$

To overcome some of the limitations of the co-occurrence approach, WikiGene is also used to complement LitMiner. WikiGene is a Wiki-based curation tool for expert users to verify and improve gene annotation data. Figure 2.3 shows the output from LitMiner using “colon cancer” as the query.


GSF - Forschungszentrum für Umwelt und Gesundheit
gsf
LitMiner
Home | Genes | Compounds | Diseases and Phenotypes | Tissues and Organs
Contact
Apply Filters

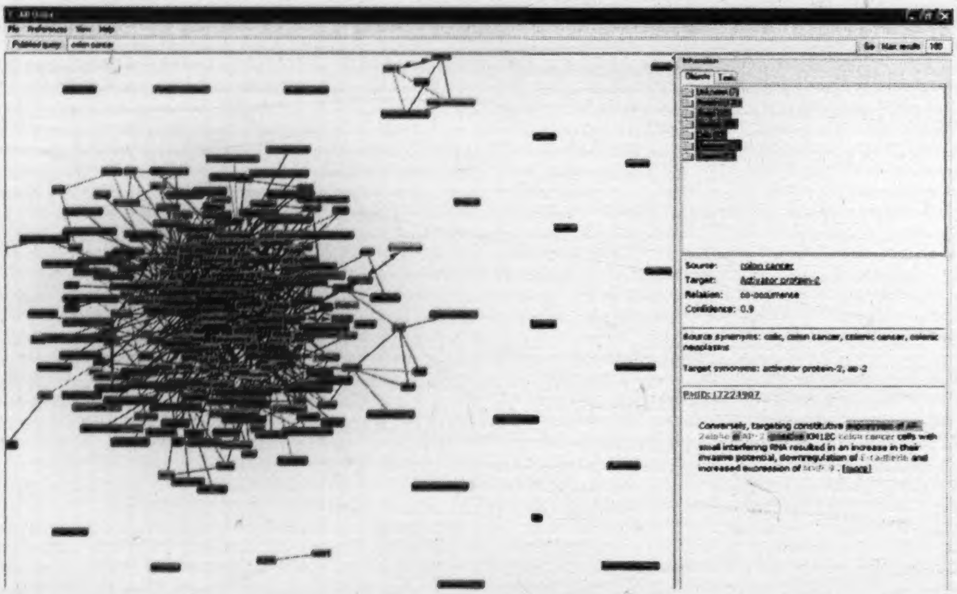
Genes co-annotated with disease/phenotype: "Colon cancer" (10800 articles)

Gene	Species	Locus	Is TF Filter: ● all ○ TF	Articles Filter > 10	Number of co-annotated articles Filter > 10	Overrepresentation score Filter > 3	Apply Filters
Ptgri	*Mus musculus*	2E1		4	2	596	Co-annotated tissues/organs
AREG	*Homo sapiens*	4q13.3		5	2	477	Co-annotated tissues/organs
AXIN1	*Homo sapiens*	16p13.3		9	3	397	Co-annotated tissues/organs
Gpa33	*Mus musculus*	1H2		17	5	350	Co-annotated tissues/organs
FACL4	*Homo sapiens*	Xq23		7	2	340	Co-annotated tissues/organs
Facl4	*Mus musculus*	XF1		7	2	340	Co-annotated tissues/organs
Q9WYS2	*Rattus norvegicus*	15p14		11	3	325	Co-annotated tissues/organs
Pms2	*Mus musculus*	5G2		55	13	282	Co-annotated tissues/organs
PMS2	*Homo sapiens*	7p22.1		55	13	282	Co-annotated tissues/organs
Drg1	*Mus musculus*	11A1		9	2	265	Co-annotated tissues/organs
DRG1	*Homo sapiens*	22q12.3		9	2	265	Co-annotated tissues/organs
KLF4	*Homo sapiens*	9q31.3	X	17	3	210	Co-annotated tissues/organs
Klf4	*Mus musculus*	4B3		24	4	199	Co-annotated tissues/organs

Figure 2.3: LitMiner's Disease – Gene output for colon cancer.

As seen in Figure 2.3, LitMiner returns a list of genes for colon cancer ranked by the over-representation score. Users can also select filters to limit the number of results displayed by choosing minimum a number of articles, co-annotated articles or a specific over-representation score. As with MedGene, the only evidence that LitMiner provides for associations is a statistical score (the over-representation score). No other supporting sentences or "facts" are provided. Furthermore, LitMiner does not provide any hyperlinks to the abstracts used to calculate the over-representation score. This makes it very difficult for users to determine how and why disease X is associated with gene Y. Unless the users go back to PubMed and search for disease X and gene Y explicitly, there is no effective way of determining whether the associations provided by LitMiner are true or not. One might ask if an overrepresentation score of 160 is high enough to be a true association. Is an overrepresentation score of 50 too low to be a true association? No evaluation of LitMiner's performance was reported; however, LitMiner has been used by others for a pathway-oriented analysis of multiple gene lists [21].

ALIBABA is a text mining tool that extracts associations between cells, diseases, drugs, proteins, species and tissues [12]. ALIBABA provides a graphical interface in an attempt to represent the extracted information as a graph. Figure 2.4 shows an example output from ALIBABA with the query "colon cancer".



Source: colon cancer
Target: Activator protein-2
Relation: co-occurrence
Confidence: 0.9

Source synonyms: colic, colon cancer, colonic cancer, colonic neoplasms

Target synonyms: activator protein-2, ap-2

PMID: 17224907

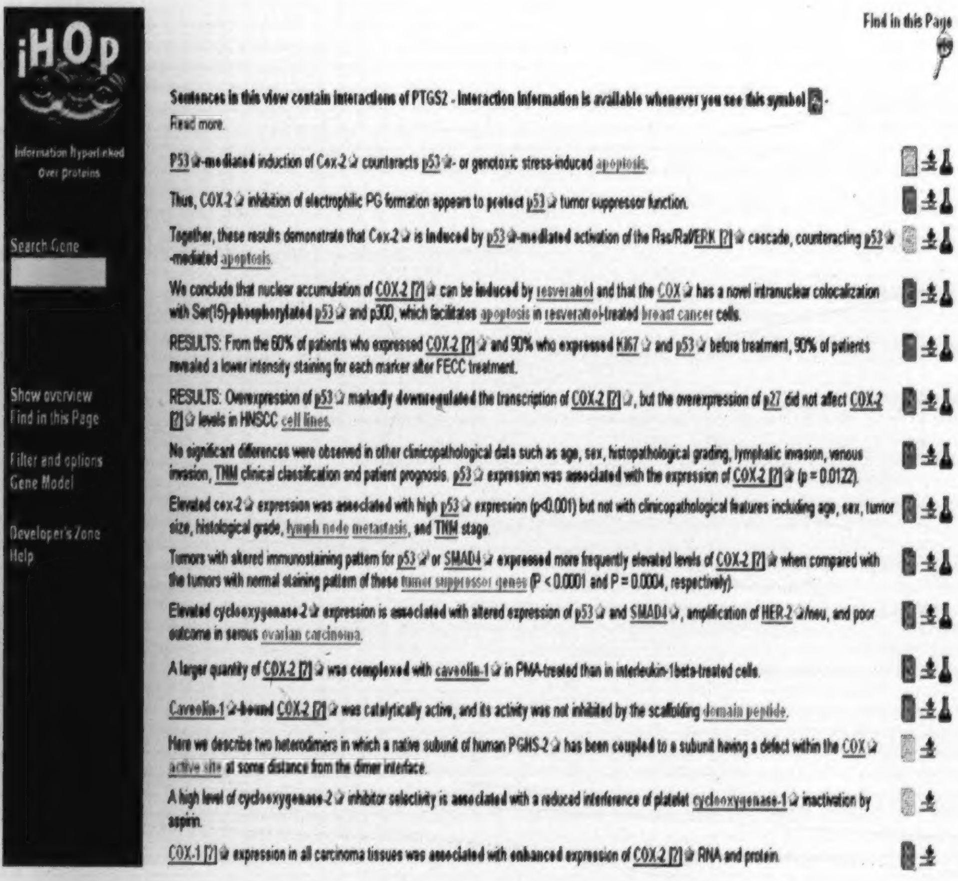
Conversely, targeting constitutive expression of AP-2alpha (AP-2) in KM12C colon cancer cells with small interfering RNA resulted in an increase in their invasive potential, downregulation of E-cadherin and increased expression of MMP-9. [more]

Figure 2.4: Output from ALIBABA using colon cancer as the query and maximum number of abstracts is set as 100.

When ALIBABA is queried with a subject yielding a large number of associations (such as colon cancer), the graphical clustering becomes a poor method for visualizing or navigating data due to an over-abundance of overlapping information. Figure 2.4 shows the query result of only 100 abstracts. Currently there are more than 55,000 "colon cancer" abstracts in PubMed. It is difficult enough to find just the "colon cancer" entity in Figure 2.4 considering we are ultimately interested in the other bio-entities (cells, diseases, drugs, proteins, species and tissues) that colon cancer is related to. This defeats the purpose of using a visualization tool to abstract away details into a more understandable form. Visualizing and navigating a biological network typically requires concrete biological knowledge and an ability to recognize the linkages between bio-entities through careful examination of the data. But with ALIBABA, the process of visualizing a biological network becomes almost impossible as the tightly clustered graphical representation overburdens the user. Furthermore, the highly variable accuracy of the linkages between bio-entities provided by ALIBABA could leave users with an inaccurate impression of a biological network. With a smaller number of abstracts, ALIBABA is more effective (the default maximum number of abstracts is 20). However, with the amount of information available in PubMed, the presentation of results in the form of hyperlinked texts or tables is generally more effective.

ALIBABA uses a dictionary-based approach for recognizing biomedical objects in abstracts and uses a language pattern matching system to help the association extraction process. ALIBABA's pattern matching algorithm also provides a confidence score for each association indicating the quality of the match between the sentence and a pattern. ALIBABA's protein-protein interaction extraction module achieves an F1-measure of 61% (precision 75% and recall 52%), as evaluated on the SPIES corpus of protein-protein interaction [22]. In addition to the pattern matching algorithm, ALIBABA also uses the co-occurrence approach to ensure higher recall. As opposed to the abstract level co-occurrence approach seen in MedGene and LitMiner, ALIBABA uses word co-occurrence at the sentence level. In general, ALIBABA is better than MedGene and LitMiner at conveying how and why it is asserting that X is associated with Y by providing a limited number of key sentences in addition to a confidence score. While the confidence score is another means of showing the strength of associations between X and Y, it still lacks biological meaning. Likewise choosing the right cut-off confidence score for automatic information extraction remains an open question for this particular tool.

IHOP stands for Information Hyperlinked Over Proteins [11]. This biomedical text mining system, as its name suggests, is designed to support only gene/protein associations and interactions. Unlike most other tools, IHOP allows searches for genes of multiple organisms including Homo sapiens. As with ALIBABA, IHOP uses a dictionary-based approach for identifying the genes/proteins in the abstract text. The strength of IHOP is that it applies a list of heuristics to rank the gene synonyms and the key sentences to ensure that the extracted information is of high quality. For example, a short gene name is ranked lower than a longer gene name and a short sentence is ranked higher than a longer sentence. In addition, IHOP also searches for MeSH (Medical Subject Heading) terms and gene-verb-gene patterns inside sentences to provide information on the genes as well as further improve the quality of extracted information. Figure 2.5 shows an example output from IHOP with the query gene "COX-2".



Sentences in this view contain interactions of PTGS2 - Interaction information is available whenever you see this symbol -
Read more.

P53-mediated induction of Cox-2 counteracts p53- or genotoxic stress-induced apoptosis.

Thus, COX-2 inhibition of electrophilic PG formation appears to protect p53 tumor suppressor function.

Together, these results demonstrate that Cox-2 is induced by p53-mediated activation of the Ras/Raf/ERK [?] cascade, counteracting p53 -mediated apoptosis.

We conclude that nuclear accumulation of COX-2 [?] can be induced by resveratrol and that the COX has a novel intranuclear colocalization with Ser(15)-phosphorylated p53 and p300, which facilitates apoptosis in resveratrol-treated breast cancer cells.

RESULTS: From the 60% of patients who expressed COX-2 [?] and 90% who expressed Ki67 and p53 before treatment, 90% of patients revealed a lower intensity staining for each marker after FECC treatment.

RESULTS: Overexpression of p53 markedly downregulated the transcription of COX-2 [?], but the overexpression of p27 did not affect COX-2 [?] levels in HNSCC cell lines.

No significant differences were observed in other clinicopathological data such as age, sex, histopathological grading, lymphatic invasion, venous invasion, TNM clinical classification and patient prognosis. p53 expression was associated with the expression of COX-2 [?] (p = 0.0122).

Elevated cox-2 expression was associated with high p53 expression (p<0.001) but not with clinicopathological features including age, sex, tumor size, histological grade, lymph node metastasis, and TNM stage.

Tumors with altered immunostaining pattern for p53 or SMAD4 expressed more frequently elevated levels of COX-2 [?] when compared with the tumors with normal staining pattern of these tumor suppressor genes (P < 0.0001 and P = 0.0004, respectively).

Elevated cyclooxygenase-2 expression is associated with altered expression of p53 and SMAD4, amplification of HER-2/neu, and poor outcome in serous ovarian carcinoma.

A larger quantity of COX-2 [?] was complexed with caveolin-1 in PMA-treated than in interleukin-1beta-treated cells.

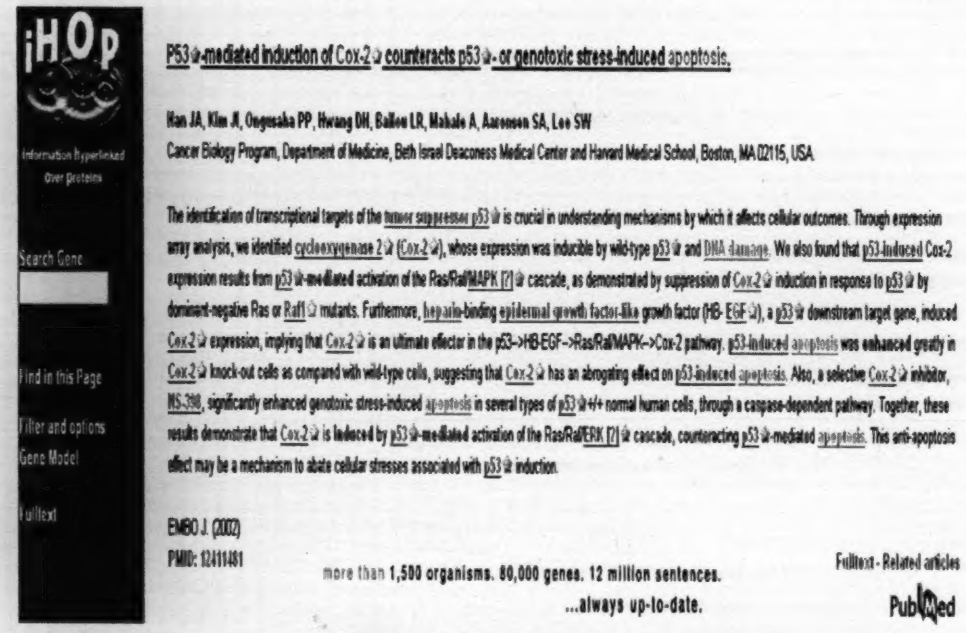
Caveolin-1-bound COX-2 [?] was catalytically active, and its activity was not inhibited by the scaffolding domain peptide.

Here we describe two heterodimers in which a native subunit of human PGHS-2 has been coupled to a subunit having a defect within the COX active site at some distance from the dimer interface.

A high level of cyclooxygenase-2 inhibitor selectivity is associated with a reduced interference of platelet cyclooxygenase-1 inactivation by aspirin.

COX-1 [?] expression in all carcinoma tissues was associated with enhanced expression of COX-2 [?] RNA and protein.

Figure 2.5: Output from IHOP using COX-2 as the query.



P53-mediated induction of Cox-2 counteracts p53- or genotoxic stress-induced apoptosis.

Han JA, Kim JI, Ongusaha PP, Hwang DH, Ballou LR, Mahale A, Aaronson SA, Lee SW
Cancer Biology Program, Department of Medicine, Beth Israel Deaconess Medical Center and Harvard Medical School, Boston, MA 02115, USA

The identification of transcriptional targets of the tumor suppressor p53 is crucial in understanding mechanisms by which it affects cellular outcomes. Through expression array analysis, we identified cyclooxygenase 2 (Cox-2), whose expression was inducible by wild-type p53 and DNA damage. We also found that p53-induced Cox-2 expression results from p53-mediated activation of the Ras/Raf/MAPK [?] cascade, as demonstrated by suppression of Cox-2 induction in response to p53 by dominant-negative Ras or Raf1 mutants. Furthermore, heparin-binding epidermal growth factor-like growth factor (HB- EGF), a p53 downstream target gene, induced Cox-2 expression, implying that Cox-2 is an ultimate effector in the p53-->HB-EGF-->Ras/Raf/MAPK-->Cox-2 pathway. p53-induced apoptosis was enhanced greatly in Cox-2 knock-out cells as compared with wild-type cells, suggesting that Cox-2 has an abrogating effect on p53-induced apoptosis. Also, a selective Cox-2 inhibitor, NS-398, significantly enhanced genotoxic stress-induced apoptosis in several types of p53+/+ normal human cells, through a caspase-dependent pathway. Together, these results demonstrate that Cox-2 is induced by p53-mediated activation of the Ras/Raf/ERK [?] cascade, counteracting p53-mediated apoptosis. This anti-apoptosis effect may be a mechanism to abate cellular stresses associated with p53 induction.

EMBO J. (2002)
PMID: 12411481

Fulltext - Related articles

more than 1,500 organisms. 80,000 genes. 12 million sentences.
...always up-to-date.
PubMed

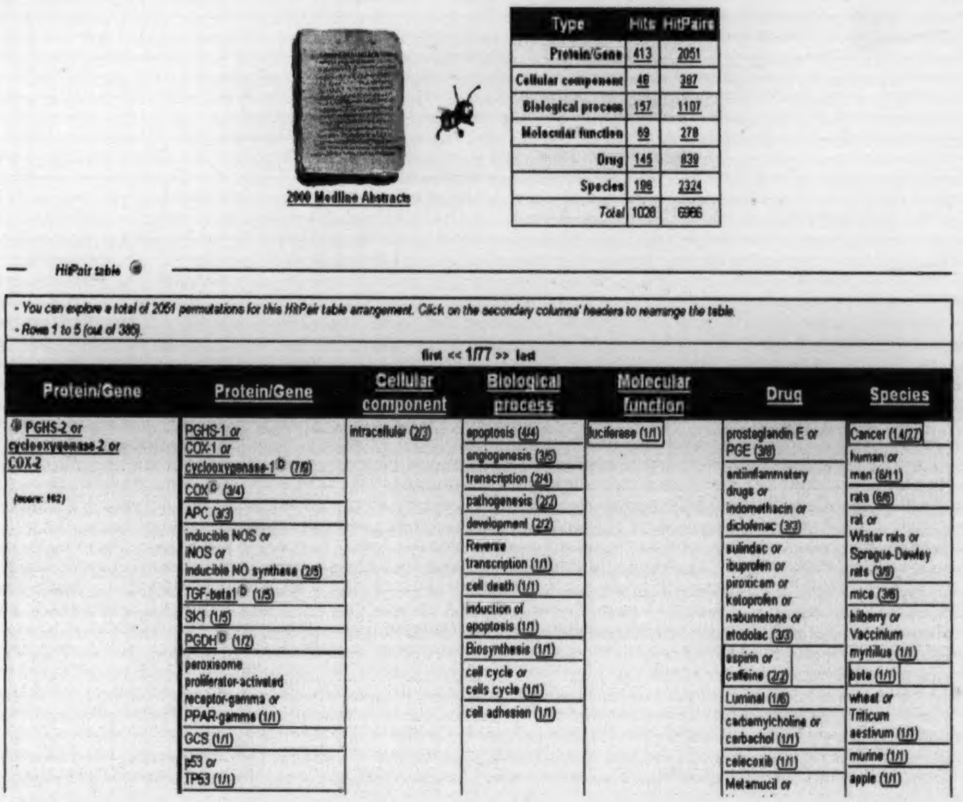
Figure 2.6: An IHOP abstract used to find COX-2 gene interactions.

IHOP displays all the key sentences (sentences that mention two or more genes or sentences that mention a single gene together with one or more MeSH terms) that it found for a given query. This allows users to read all the key sentences themselves. The abstracts that IHOP used to extract information are also available for users to read (as shown in Figure 2.6). This makes IHOP an excellent example of a tool that provides the information needed by biologists to understand the "how" and "why" of an association between X and Y.

The authors of IHOP evaluated IHOP's ability to identify the proper gene/protein names (gene synonym identification) inside abstract texts. For gene synonym identification, IHOP's performance in terms of f-measure range from 70% to 91% depending on the organism. The variation in performance comes from the fact that in organisms such as yeast and C. elegans, the gene naming convention and gene name usage in the literature is well structured whereas for organisms such as mouse and human, the gene naming conventions and gene name usage in the literature are far more varied. No evaluations of the performance of IHOP's protein-protein interaction extraction were available. The authors of IHOP describe it as a system that allows researchers to easily move or "hop" between sentences taken directly from their source abstracts. This allows users to retain control over the reliability of the information they obtain. IHOP focuses on allowing human experts to navigate through the scientific literature and to gather relevant information themselves. This is because no automatic information extraction system can achieve a comparable level of precision without a significant loss of recall. The primary limitations of IHOP lie in the fact that it is designed only to support gene/protein interaction searches and the fact that it provides a relatively limited overview of all the possible gene interactions. For example, it is hard to tell in Figure 2.5 how many gene interaction partners COX-2 has. It is also difficult to assess the strength of association between COX-2 and its gene interaction partners. In particular, IHOP does not explicitly specify a score for association strength to support the association of X with Y.

EBIMed is the latest text mining tool from the European Bioinformatics Institute [13]. EBIMed searches for protein/gene, cellular compartment, biological process, molecular function, drug, and species names inside abstracts. Like ALIBABA and IHOP, EBIMed uses a dictionary-based approach for identifying the biomedical objects in PubMed abstracts. EBIMed also uses sentence-level word co-occurrence to evaluate sentence information. For example, if protein X and protein Y co-occur in one sentence, the score of protein X increases by one. If protein X and drug W co-occur in one sentence, the score for protein X increases by one again. The total score of protein X depends on the number of sentences in which protein X co-occurs with another key term. The initial keyword query is independent of EBIMed's analysis. This means that the query can be anything (i.e. not a gene) as opposed to IHOP, which limits the initial keyword query to gene names only. To evaluate EBIMed's performance, the authors carried out four different analyses using the gene *Wnt* as the initial keyword query. In one analysis, where EBIMed's ability to identify protein names in abstract texts was evaluated, they reported >90% precision. In another analysis, the authors selected the first 20 proteins in alphabetical order identified by EBIMed using the query *Wnt* and evaluated all 94 extracted sentences for the 20 proteins that contained protein pairs. This analysis was aimed at evaluating EBIMed's ability to identify protein-protein interaction. This assessment showed that 40% of the sentences were reporting on valid protein-protein interactions. In a similar analysis for the same query but for drug-protein relations, they reported 50% of the 118 sentences containing drug-protein pairs were reporting meaningful drug-protein relations. In the final analysis, the authors evaluated EBIMed's "coverage of relation identification". For this assessment, they retrieved all sentences that contained protein pairs in the *Wnt* query and compared these results to the protein pairs found in the Wnt pathway as described in the Kyoto Encyclopedia of Genes and Genomes [(KEGG) (www.genome.jp/kegg)] and in the Signal Transduction Knowledge Environment [(STKE) (stke.sciencemag.org)]. In total, EBIMed identified 74 protein-protein pairs out of the total of 108 pairs described either in the KEGG or the STKE Wnt pathway. Using a gold standard set of data from manually curated databases as a basis for comparison is a means to test the recall of a text mining system. The assessment showed that EBIMed retrieved 68.5% of the protein-protein pairs in the Wnt pathway. If data from manually curated databases were integrated into EBIMed, then EBIMed's recall would obviously be better. Text mining systems and manually curated databases can complement each other to ensure higher precision and recall. However, EBIMed lacks the feature of integrating manually curated databases into its text mining system.

In summary, the authors of EBIMed described their system as a tool that leads to better access to key statements in biomedical text than PubMed because the user generally reads relevant sentences rather than complete abstracts. In addition, EBIMed's results tables allows users to get an overview on a multitude of relations spread over many abstracts, thereby supporting a wide variety of use scenarios. Figure 2.7 shows an example output from an EBIMed query where colon cancer was used as the query. It is useful to examine this output in more detail to appreciate the strengths and weaknesses of EBIMed.

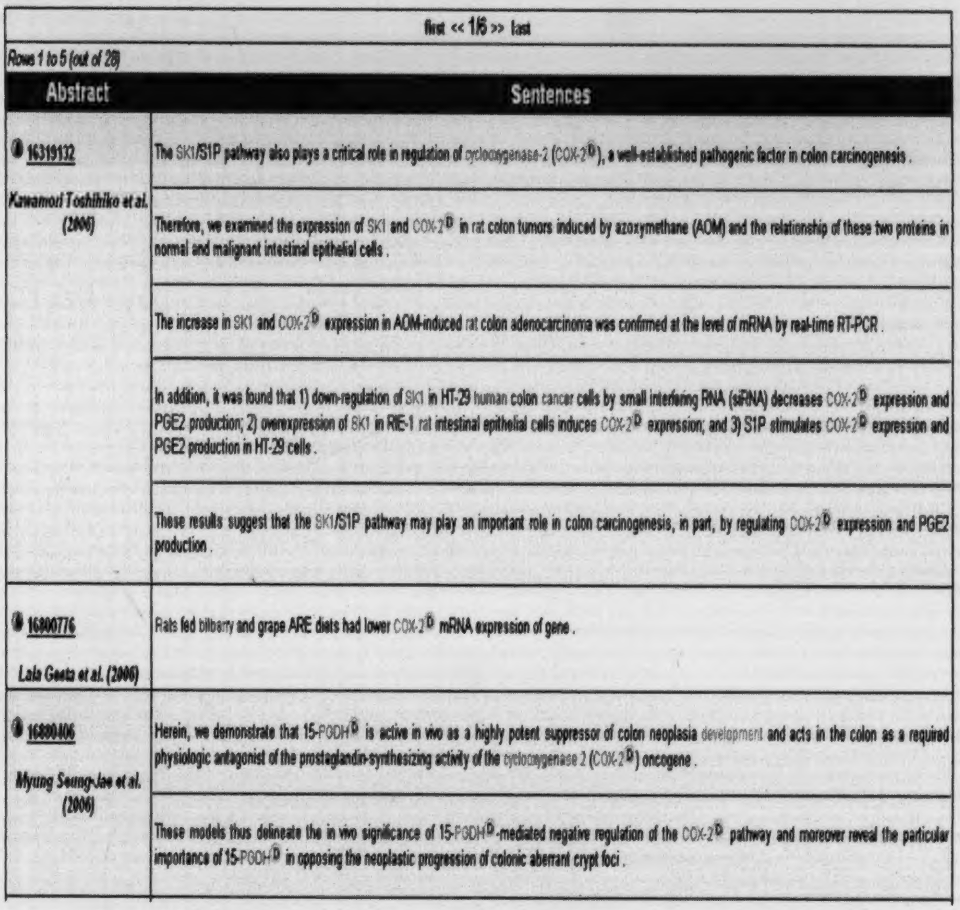


Type	Hits	HitPairs
Protein/Gene	413	2051
Cellular component	48	387
Biological process	157	1107
Molecular function	69	278
Drug	145	839
Species	196	2324
Total	1028	6986

- *You can explore a total of 2051 permutations for this HitPair table arrangement. Click on the secondary columns' headers to rearrange the table.*
- *Rows 1 to 5 (out of 385).*

Protein/Gene	Protein/Gene	Cellular component	Biological process	Molecular function	Drug	Species
PGHS-2 or cyclooxygenase-2 or COX-2 (more: 162)	PGHS-1 or COX-1 or cyclooxygenase-1 (7/9) COX (3/4) APC (3/3) inducible NOS or iNOS or inducible NO synthase (2/5) TGF-beta1 (1/5) SK1 (1/5) PGDH (1/2) peroxisome proliferator-activated receptor-gamma or PPAR-gamma (1/1) GCS (1/1) p53 or TP53 (1/1)	intracellular (2/3)	apoptosis (4/4) angiogenesis (3/5) transcription (2/4) pathogenesis (2/2) development (2/2) Reverse transcription (1/1) cell death (1/1) induction of apoptosis (1/1) Biosynthesis (1/1) cell cycle or cells cycle (1/1) cell adhesion (1/1)	luciferase (1/1)	prostaglandin E or PGE (3/8) antiinflammatory drugs or indomethacin or diclofenac (3/3) sulindac or ibuprofen or piroxicam or ketoprofen or nabumetone or etodolac (3/3) aspirin or caffeine (2/2) Luminal (1/6) carbamylcholine or carbachol (1/1) celecoxib (1/1) Metamucil or	Cancer (14/27) human or man (6/11) rats (6/6) rat or Wistar rats or Sprague-Dawley rats (3/9) mice (3/6) bilberry or Vaccinium myrtillus (1/1) beta (1/1) wheat or Triticum aestivum (1/1) murine (1/1) apple (1/1)

Figure 2.7: Output from EBIMed using "colon cancer" as the query, searching for human genes and maximum number of abstracts is set to 2000.



first << 1/6 >> last

Rows 1 to 5 (out of 28)

Abstract	Sentences
● 16319132 *Kawamori Toshihiko et al. (2006)*	The SK1/S1P pathway also plays a critical role in regulation of cyclooxygenase-2 (COX-2Ⓓ), a well-established pathogenic factor in colon carcinogenesis .
	Therefore, we examined the expression of SK1 and COX-2Ⓓ in rat colon tumors induced by azoxymethane (AOM) and the relationship of these two proteins in normal and malignant intestinal epithelial cells .
	The increase in SK1 and COX-2Ⓓ expression in AOM-induced rat colon adenocarcinoma was confirmed at the level of mRNA by real-time RT-PCR .
	In addition, it was found that 1) down-regulation of SK1 in HT-29 human colon cancer cells by small interfering RNA (siRNA) decreases COX-2Ⓓ expression and PGE2 production; 2) overexpression of SK1 in RIE-1 rat intestinal epithelial cells induces COX-2Ⓓ expression; and 3) S1P stimulates COX-2Ⓓ expression and PGE2 production in HT-29 cells .
	These results suggest that the SK1/S1P pathway may play an important role in colon carcinogenesis, in part, by regulating COX-2Ⓓ expression and PGE2 production .
● 16800776 *Lala Geeta et al. (2006)*	Rats fed bilberry and grape ARE diets had lower COX-2Ⓓ mRNA expression of gene .
● 16880406 *Myung Seung-Jae et al. (2006)*	Herein, we demonstrate that 15-PGDHⒹ is active in vivo as a highly potent suppressor of colon neoplasia development and acts in the colon as a required physiologic antagonist of the prostaglandin-synthesizing activity of the cyclooxygenase 2 (COX-2Ⓓ) oncogene .
	These models thus delineate the in vivo significance of 15-PGDHⒹ-mediated negative regulation of the COX-2Ⓓ pathway and moreover reveal the particular importance of 15-PGDHⒹ in opposing the neoplastic progression of colonic aberrant crypt foci .

Figure 2.8: Key sentences found by EBIMed for associations between colon cancer and COX-2.

As seen in Figure 2.7, COX-2 has a EBIMed score of 102 which may be of some biological significance. The score means that in the abstracts that mention colon cancer, COX-2 is an important gene since it co-occurs with other key terms most frequently (102 abstracts mention COX-2 plus its synonyms along with one or more key terms from protein/gene, cellular compartment, biological process, molecular function, drug or species). However, EBIMed seems to have relatively loose definitions for some of their key terms. In Figure 2.7, for example, one can ask if "luciferase" is a function or not. Furthermore, under the Drug column, "aspirin" and "caffeine" are grouped together suggesting that they are related but the relationship is not obvious. Also, "Cancer" and "beta" are identified as a species and obviously this is not true. EBIMed's somewhat questionable definitions of key terms can obviously affect its scores and what EBIMed considers to be relevant sentences. Looking at Figure 2.8, one can see that many of the key sentences do provide explanations of how colon cancer and COX-2 are associated. However, the sentence "Rats fed bilberry and grape ARE diets had lower COX-2 mRNA expression of gene" (*) is also considered as a relevant sentence because COX-2 co-occurred with the key species terms "Rats" and "bilberry". Since the initial keyword query was "colon cancer", the sentence (*) should be irrelevant because the sentence offers no evidence of association between colon cancer and COX-2. This sentence is included because the initial keyword query is independent of EBIMed's textual analysis. Therefore key sentences such as this one may or may not provide evidence of association between colon cancer and COX-2. In other words, the keyword query "colon cancer" is only used to retrieve abstracts from PubMed and is not considered in the key sentence and relevant scoring analysis of the abstracts. As a result of EBIMed's relaxed definitions of key terms and the questionable method used in determining sentence relevancy, using EBIMed as an automatic information extraction tool can be challenging. In particular EBIMed's score may or may not indicate the biological significance of an association between the initial keyword query and EBIMed's final results.

Chapter 3

3. PolySearch System Overview

3.1 Overview

PolySearch, as the name suggests, is a tool that supports multiple ("poly") types of biomedical text searches from multiple ("poly") types of databases. It is also designed to facilitate the search, retrieval and compilation of disease-associated human "poly"morphisms (SNPs). PolySearch exploits recent advances in text mining along with the ready availability of diverse biomedical databases and biomedical thesauruses to permit a wide variety of complex or expansive text searches over many biomedical domains. PolySearch consists of seven basic components: 1) a web-based user interface for constructing queries; 2) a collection of internal and external biomedical databases; 3) a collection of biomedical synonyms (custom thesauruses); 4) a general text search engine for extracting data from heterogeneous databases; 5) a schema for selecting, ranking and integrating content; 6) a display tool for displaying and synopsizing results and 7) a PCR primer designing tool to facilitate SNP and mutation studies. An outline of PolySearch's general design is given in Figure 3.1.

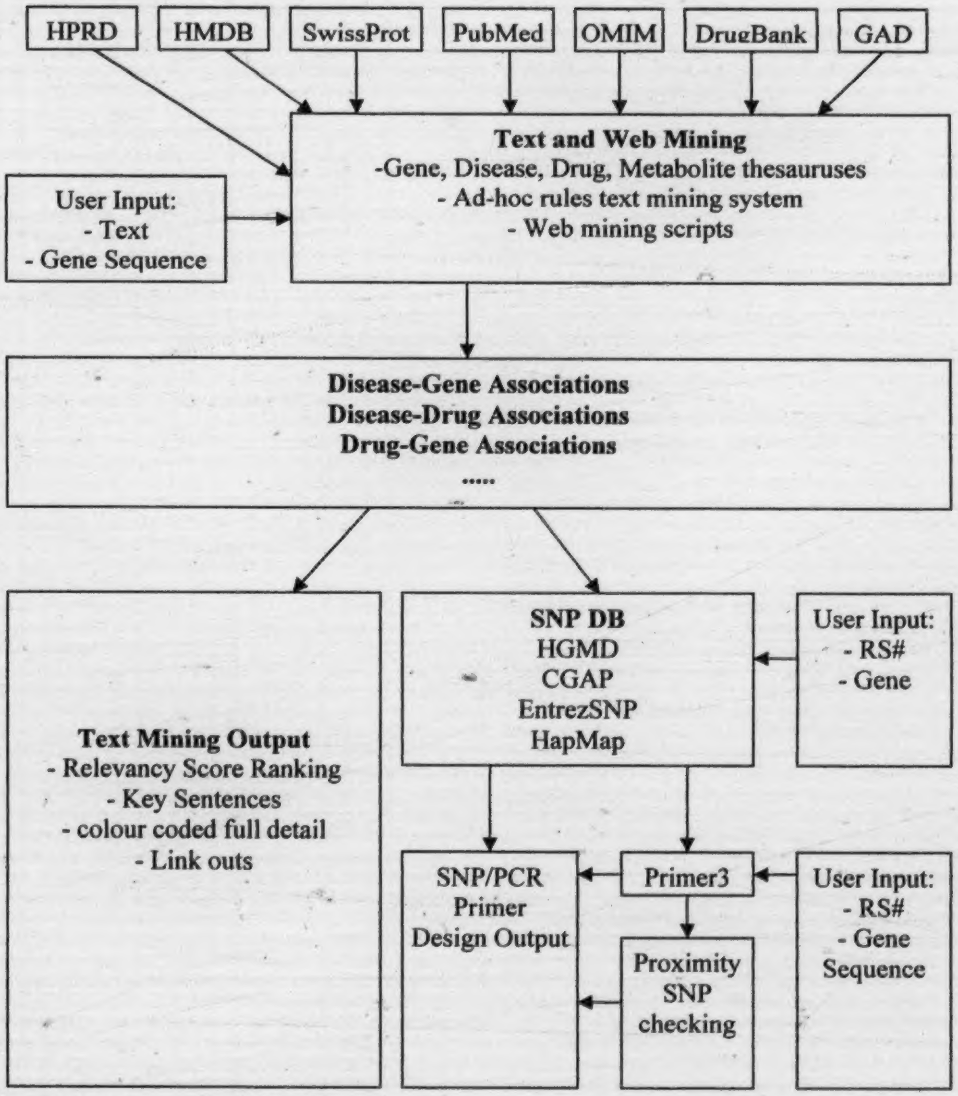
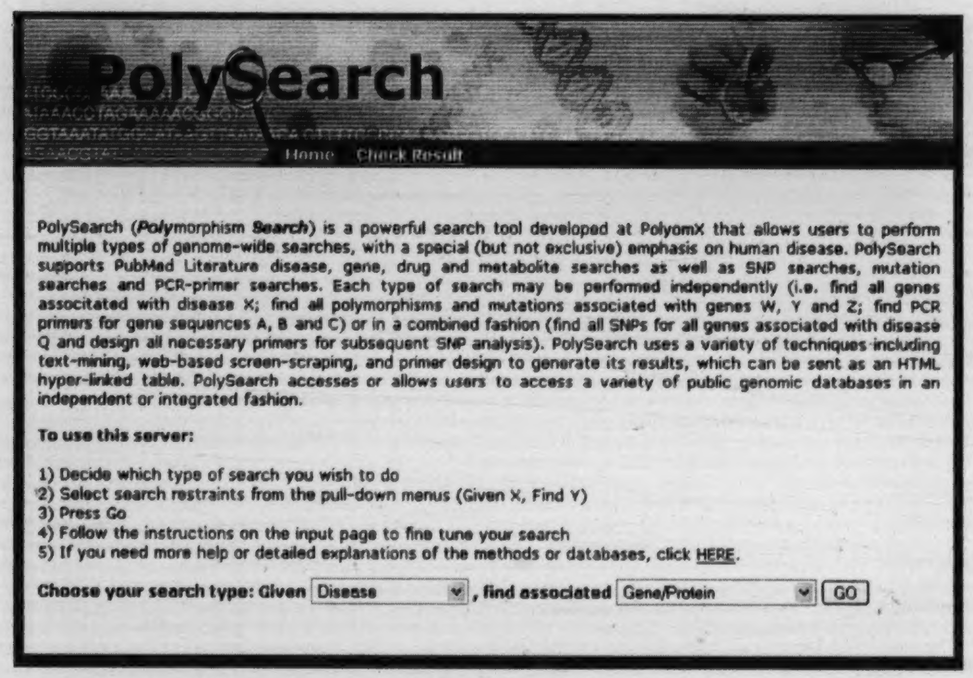

HPRD
HMDB
SwissProt
PubMed
OMIM
DrugBank
GAD
Text and Web Mining
-Gene, Disease, Drug, Metabolite thesauruses
- Ad-hoc rules text mining system
- Web mining scripts
User Input:
- Text
- Gene Sequence
Disease-Gene Associations
Disease-Drug Associations
Drug-Gene Associations
.....
Text Mining Output
- Relevancy Score Ranking
- Key Sentences
- colour coded full detail
- Link outs
SNP DB
HGMD
CGAP
EntrezSNP
HapMap
User Input:
- RS#
- Gene
SNP/PCR
Primer
Design Output
Primer3
Proximity
SNP
checking
User Input:
- RS#
- Gene
Sequence

Figure 3.1: PolySearch system overview showing the resources that PolySearch uses and the features found in PolySearch.

3.2 Query Interface

PolySearch's query interface was written in Perl and uses a series of text boxes and pull-down menus to facilitate query construction. A Screen shot of the query interface is shown in Figure 3.2.



PolySearch (***Poly***morphism ***Search***) is a powerful search tool developed at PolyomX that allows users to perform multiple types of genome-wide searches, with a special (but not exclusive) emphasis on human disease. PolySearch supports PubMed Literature disease, gene, drug and metabolite searches as well as SNP searches, mutation searches and PCR-primer searches. Each type of search may be performed independently (i.e. find all genes associtated with disease X; find all polymorphisms and mutations associated with genes W, Y and Z; find PCR primers for gene sequences A, B and C) or in a combined fashion (find all SNPs for all genes associated with disease Q and design all necessary primers for subsequent SNP analysis). PolySearch uses a variety of techniques including text-mining, web-based screen-scraping, and primer design to generate its results, which can be sent as an HTML hyper-linked table. PolySearch accesses or allows users to access a variety of public genomic databases in an independent or integrated fashion.

To use this server:

1) Decide which type of search you wish to do
2) Select search restraints from the pull-down menus (Given X, Find Y)
3) Press Go
4) Follow the instructions on the input page to fine tune your search
5) If you need more help or detailed explanations of the methods or databases, click HERE.

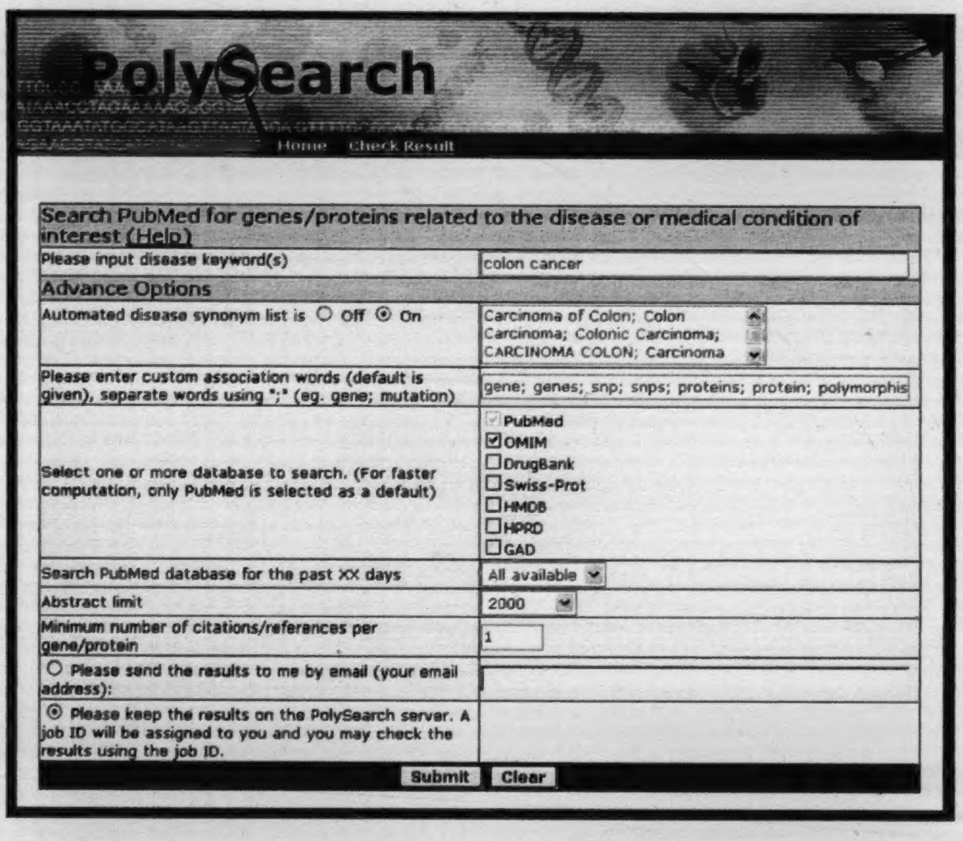
Choose your search type: Given Disease **, find associated** Gene/Protein

Figure 3.2: PolySearch's homepage where users can select the different "given X find associated Y's" queries.

The basic structure of almost every PolySearch query is "given a single X find all associated Y's", where X can be any human disease name, gene/protein name, drug name, metabolite name, SNP, gene/protein sequence or user-provided text word and Y can be any one of all human diseases, genes/proteins, drugs, metabolites, tissues, organs, subcellular localizations, SNPs, PCR primers or user supplied text words. In each case the "X" and "Y" words can correspond to either a common name or synonyms. Table 3.1 provides a more detailed listing of all allowed "basic" queries in PolySearch. Once the general query is constructed and submitted the user is presented with a second page (the query refinement page, Figure 3.3) that allows further refinement of the query, including the selection of association words, databases, query-word synonyms and display options.

		Given						
		Disease	**Gene/ Protein**	**Drug**	**Metabolite**	**Text word**	**SNP (RS #)**	**Gene/ Protein Sequence**
Find	**Disease**	✓	✓	✓	✓	✓		✓
	Gene/ Protein	✓	✓	✓	✓	✓	✓	✓
	Drug	✓	✓	✓	✓	✓		✓
	Metabolite	✓	✓	✓		✓		✓
	Tissue	✓	✓	✓	✓	✓		✓
	Organ	✓	✓	✓	✓	✓		✓
	Subcellular Localization	✓	✓	✓	✓	✓		✓
	Text word					✓		
	SNP		✓					✓
	PCR Primers						✓	✓

Table 3.1: A detailed listing of all allowed "basic" queries in PolySearch.



Search PubMed for genes/proteins related to the disease or medical condition of interest (Help)	
Please input disease keyword(s)	colon cancer
Advance Options	
Automated disease synonym list is ○ Off ● On	Carcinoma of Colon; Colon Carcinoma; Colonic Carcinoma; CARCINOMA COLON; Carcinoma
Please enter custom association words (default is given), separate words using ";" (eg. gene; mutation)	gene; genes; snp; snps; proteins; protein; polymorphis
Select one or more database to search. (For faster computation, only PubMed is selected as a default)	☑ PubMed ☑ OMIM ☐ DrugBank ☐ Swiss-Prot ☐ HMDB ☐ HPRD ☐ GAD
Search PubMed database for the past XX days	All available
Abstract limit	2000
Minimum number of citations/references per gene/protein	1
○ Please send the results to me by email (your email address):	
● Please keep the results on the PolySearch server. A job ID will be assigned to you and you may check the results using the job ID.	

Figure 3.3: The query refinement page for "Given Disease Find Associated Gene".

The majority of PolySearch's queries depend on text searches through PubMed abstracts. To facilitate PubMed searching we use the E-utilities application programming interface (API) from NCBI that allows abstracts to be batch-downloaded from the PubMed website [23]. The downloaded abstracts are then searched on the PolySearch server (i.e. locally) for key sentences through PolySearch's own text mining tools. By default, PolySearch uses a set of pre-defined association words to make its searches through the extracted abstracts more specific. To understand the need for these association words, consider the type and quantity of information that is available for a well-studied disease such as "breast cancer". The 162,000 abstracts in PubMed containing the words "breast cancer" include genetic information about breast cancer, clinical symptoms of breast cancer, surgical procedures about breast cancer, socio-economic data about breast cancer, psycho-social data about breast cancer along with many other aspects concerning this disease. Searching through such a large collection of heterogeneous abstracts would obviously lead to heterogeneous results concerning the molecular etiology of the disease. Limiting the results to include only sentences that contain association words that are concerned with the molecular or genetic aspects of breast cancer would likely return a more properly ranked homogenous result. Obviously not all users are interested in molecular or genetic aspects of certain conditions so PolySearch also allows users to enter their own association words in order to customize the scope or extent of their queries. This customizable word association is a particularly unique feature for PolySearch. The association words also play a key role in ranking the results. An illustration of how to choose the association words is shown in the next chapter.

Through its query refinement page (Figure 3.3), PolySearch also allows users to add or include synonyms to their original query words (i.e. query synonym expansion). Normally, PubMed, through its large collection of MeSH terms and cross-indexing, automatically generates synonyms prior to performing its searches. For instance, any query to PubMed that includes the word "yeast" is automatically modified to include "saccharomyces" in the search term.

However, not all MeSH terms have synonyms nor are these synonyms very complete. Therefore, PolySearch uses its own thesauruses to automatically append synonyms to a query word (by clicking on the option for "automated synonym list"). If the computer-generated synonyms appear inadequate, the user may further edit or add to this list. The intelligent use of synonyms (for a disease, gene name, protein name, drug or metabolite) for a query word can greatly improve the specificity and sensitivity of a given search.

From the query refinement interface users can also choose to limit their search to PubMed only, or to perform their search on some of PolySearch's other reference databases (i.e. OMIM, SwissProt, DrugBank, HMDB, etc.). Limiting PolySearch searches to the PubMed database (the default configuration) is faster and it allows users to quickly assess what may be necessary to refine their queries for better results. Once the queries are refined, users may then expand their search to include the other databases by clicking on the appropriate database checkboxes. Additionally, through the query refinement interface users can also specify: 1) how far back in time the PubMed records should be searched, 3) the number of abstracts to be searched and 3) the minimum number of PubMed citations required to be considered as a hit.

PolySearch is not limited to basic queries such as "given a single X find all associated Y's". More complex queries can be assembled using different combinations of PolySearch's basic queries. For example, if a user wanted to find a list of SNPs associated with a disease they would first find the genes associated with the disease using the "Given Disease-Find Gene" query and then from the resulting list of genes, perform multiple "Given Gene-Find SNP" queries to get the complete list of SNPs.

Once a query is submitted, PolySearch will retrieve PubMed abstracts and fields from the appropriate databases relevant to the query (see below for more details). From these abstracts or database synopses/descriptions, PolySearch further parses them into their component sentences. This sentence parsing improves the automated extraction of the most informative phrases or facts [24]. PolySearch also uses a variety of ad hoc rules to search for the query words, association words and words from PolySearch's thesauruses in order to rank the sentences in terms of relevancy.

3.3 PolySearch's Databases

One of the more unique features of PolySearch is its integration of multiple databases containing both text and sequence data. Currently PolySearch can search and extract data from more than a dozen biomedical databases including PubMed, OMIM [25], SwissProt [4], DrugBank [6], the Human Metabolome Database (HMDB) [7] the Human Protein Reference Database (HPRD) [26], the Genetic Association Database (GAD) [27], HapMap [28], Entrez SNP (dbSNP) [3], CGAP SNP500cancer Database [29], and the Human Genome Mutation Database (HGMD) [5]. Many of these databases (PubMed, OMIM, etc.) are housed externally and queried through various custom CGI tools written in Perl, while others (DrugBank, HMDB and the SNP databases) are housed internally to accelerate PolySearch's query process. Below is a short description of each database.

- PubMed: A service of the U.S. National Library of Medicine that includes over 16 million abstracts and paper titles from life science journals dating back to the 1950s.

(http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?DB=pubmed)

- OMIM (Online Mendelian Inheritance in Man): A catalog of human genes and genetic disorders authored and edited by Dr. Victor A. McKusick and his colleagues at Johns Hopkins University, and developed for the World Wide Web by the NCBI (the National Center for Biotechnology Information).

(http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM)

- GAD (Genetic Association Database): An archive of human genetic association studies of complex diseases and disorders.

(http://geneticassociationdb.nih.gov/)

- SwissProt: A curated protein sequence database that strives to provide a high level of annotation (such as the description of the function of a protein, its domains structure, post-translational modifications, variants, etc.), a minimal level of redundancy and high level of integration with other databases.

(http://www.expasy.org/sprot/)

- HPRD (Human Protein Reference Database): A centralized platform to visually depict and integrate information pertaining to domain architecture, post-translational modifications, interaction networks and disease association for each protein in the human proteome.

(http://www.hprd.org/)

- DrugBank: A unique bioinformatics and cheminformatics resource that combines detailed drug (i.e. chemical, pharmacological and pharmaceutical) data with comprehensive drug target (i.e. sequence, structure, and pathway) information.

(http://redpoll.pharmacy.ualberta.ca/drugbank/)

- HMDB: A freely available electronic database containing detailed information about small molecule metabolites found in the human body.

(http://www.hmdb.ca/)

- HapMap: A freely available resource that contains information pertaining to haplotype map of the human genome. The HapMap database describes the common patterns of human DNA sequence variation.

(http://www.hapmap.org/)

- Entrez SNP (dbSNP): A central repository for both single base nucleotide substitutions (SNPs) and short deletion and insertion polymorphisms in the human genome.

(http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Snp)

- CGAP SNP500cancer Database: A part of the Cancer Genome Anatomy Project and is specifically designed to contain data on the genetic variation in genes important in cancer.

(http://snp500cancer.nci.nih.gov/home_1.cfm)

- Human Genome Mutation Database: A database comprises various types of mutation within the coding regions, splicing and regulatory regions of human nuclear genes causing inherited disease.

(http://www.hgmd.cf.ac.uk/ac/index.php)

A key challenge in working with so many databases lies with the heterogeneity of the data and the diversity of data formats. Specifically, each of PolySearch's reference databases contains information in different formats requiring customized querying and formatting tools. For instance, PubMed and OMIM contain data in the form of paragraphs and complex English sentences. On the other hand, SwissProt, DrugBank, HMDB and many of the SNP/mutation databases contain numeric data and synoptic sentences associated with various labeled fields. All of these databases have their own unique formats and peculiarities. Nevertheless, the inclusion of any of these databases in a PolySearch query can greatly add to the richness or information content of a given result. This is because not all information about a gene, disease, protein, drug or metabolite is necessarily contained in a PubMed abstract. Indeed most of this kind of information is in the body of scientific papers and in textbooks. Since many databases contain information derived from these sources, it is often true that some databases contain much more valid information than can be found in abstracts alone. As will be shown later, the inclusion of additional databases can significantly improve PolySearch's performance.

3.4 Custom Thesauruses

In addition to its own unique collection of databases and database searching tools, PolySearch also has its own very extensive lists of manually curated synonyms for human genes, human proteins, diseases, drugs, metabolites, tissues, organs and subcellular localizations. These synonym lists or custom thesauruses are critical for many of the expansive queries ("given one, find many") supported by PolySearch. They are also critical for providing the sensitivity and specificity for many single word queries (i.e. the implementation of the automated synonym feature in the query refinement page). For example, PolySearch's human gene/protein thesaurus is compiled and updated from latest releases of SwissProt [4], Entrez Gene [3], the Human Genome Organisation Gene Nomenclature Committee (HGNC) [15] and the Human Protein Reference Database (HPRD) [26]. Two of the databases (Entrez Gene and HGNC) specialize in gene names while the other two databases (Swiss-Prot and HPRD) specialize in protein names. PolySearch's gene/protein thesaurus includes both gene and protein names, gene symbols, gene/protein abbreviations as well as their known synonyms. However these integrated lists still require considerable human editing. Specifically, the list must be filtered to exclude names, symbols or synonyms that are nonsensical or less meaningful such as DKFZp686F0970, MGC20392, Hypothetical protein FLJ37794, or C1orf60. Additionally a number of gene name expansions are also preferred to improve performance. For example, IGLC1 is expanded to IGLC-1 (also vice versa) while TNF alpha is expanded to TNFalpha and TNFa. In some cases gene names or gene/protein abbreviations are not sufficiently unique or can be easily confused with common English words (e.g. TH1 like, ARE, hole) and so these are removed. Other proteins or genes appear to be absent from these lists (cyclooxygenase 1, SCPX thiolase) or they are referred to as a single entity (ribosome, alcohol dehydrogenase) even though they are composed of multiple subunits. Therefore, a list of protein complexes or protein family names was created via manual curation to ensure PolySearch is not missing common protein complex or protein family names.

PolySearch's disease thesaurus is derived from the Unified Medical Language System (UMLS) [30] which is further supplemented with manual curation. PolySearch's drug thesaurus consists of a list of drug names and synonyms from DrugBank's list of FDA approved drugs, while its metabolite thesaurus consists of a list of metabolite names and synonyms from all entries in the HMDB. To create the tissue and organ thesauruses, the tissue and organ list from LitMiner was first combined with a tissue and organ list manually derived from the tissue specificity field in Swiss-Prot. This combined tissue and organ list was parsed and edited manually to create a separate tissue thesaurus and a separate organ thesaurus. Finally, the subcellular localization thesaurus was created from the list of all possible subcellular localizations listed in HPRD. Table 3.2 shows a summary of number of names and synonyms each thesaurus has in PolySearch. Users may also provide their own thesauruses to permit more specialized searches. Details for constructing and using these custom thesauruses are provided on the PolySearch help page. (http://wishart.biology.ualberta.ca/polysearch/help/PolySearchHelp.htm).

	Number of Unique Entries	Number of Names/Synonyms
Disease	26625	75154
Gene/Protein	26388	179320
Drug	1566	24013
Metabolite	2753	28887
Tissue	955	985
Organ	104	201
Subcellular Localization	74	175

Table 3.2: Statistics for PolySearch's thesauruses.

Chapter 4

4. PolySearch Data Mining

4.1 PolySearch's Text Mining System

The goal of a text mining system is to extract meaningful information from textual data. To achieve this goal, PolySearch's text mining system brings together a collection of tools including its own extensive lists of thesauruses, access to textual data of different databases, a unique scoring scheme, and a results display that aids the data mining process. The following sections describe each component of PolySearch's text mining system in depth, beginning with the first step in the text mining process: retrieving textual data from databases and identifying the different thesaurus terms within the retrieved texts.

The database or databases that PolySearch should search depends on the type of query the user chooses to carry out, as different databases offer different specialized data. For text mining of PubMed, OMIM, DrugBank, SwissProt, HMDB, HPRD and GAD, all textual content is treated as an abstract or "pseudo-abstract". To handle these databases, the search functionality of the respective databases is used to retrieve records relevant to the query. For PubMed, the records are already in a suitable abstract form. With OMIM the fields text, description, gene function, molecular genetics, clinical features, clinical management, and biochemical features are used to identify and extract as the "pseudo-abstracts" within each OMIM entry. With SwissProt the reference and comment fields are used as the pseudo-abstracts for text processing. In particular, the comment field contains sub-fields such as function, catalytic activity, subcellular location, tissue specificity, etc. For DrugBank, the fields Indication, Pharmacology, Mechanism Of Action, Phase 1 Metabolizing Enzyme, Drug Target Names, Drug Target Gene Names, and Drug Target Synonyms are used to parse out the relevant abstracts. For HMDB, the fields Description, Associated Disorders, Metabolic Enzyme Names, Metabolic Enzyme Synonyms, and Metabolic Enzyme Gene Names are used for abstract identification. For HPRD, the fields Interactions, Diseases, Localization, and Expressions are used to construct the pseudo-abstracts. Finally, for GAD, we use the fields Gene Symbol, Gene Name, Reference Title, and Study Conclusion to parse out the relevant pseudo-abstracts.

Once the appropriate abstracts, pseudo-abstracts or paragraphs have been identified and downloaded from the relevant databases, PolySearch then proceeds to analyze them using its own text mining utilities. PolySearch begins the text mining process by parsing each abstract into individual sentences. Negative sentences such as "Experimental results have shown that disease A is not associated to gene Y." are first removed from consideration. PolySearch identifies negative or negative-result sentences by searching for one of the following negative keywords or phrases: "not observed", "no evidence", "not present", "invalid", "not validated", "not proven", "no proof", "insufficient evidence", "doubtful", "not show(n)", "unproven", "unlikely", "not likely", "not associate(d)", "unassociated", "not see(n)", "not linked", "unsupported", "exclude(d)", "not include", "negative", and "exclusion". After the removal of negative sentences, the remaining sentences are tokenized into individual words or word clusters and so-called "stop words" are removed. Stop words are very commonly used words such as conjunctions, prepositions and articles that contribute little to relevancy or content. The stop-word list for PolySearch also contains some medical stop words such as "cell", "clinical", "effect", "growth", "health", "human", "medical", "medicine", "patient", etc. After the removal of the stop words, the remaining words or tokens are used to check against the query words (and their corresponding synonyms) to identify and rank relevant sentences. In general if an abstract or a sentence in the abstract contains the query word(s) along with one or more matching words to the appropriate disease, gene/protein, drug or metabolite then the abstract or sentence is marked for further processing. Once an abstract or pseudo abstract is identified as being relevant, PolySearch also keeps a record of which database they were derived from. In the case of PubMed abstracts, information about their PubMed IDs, the journals that they were published in, the time of publication and the publication type are recorded.

Before trying to find possible associations in the text, it is important first to have consistent and correct identification of appropriate query/thesaurus terms and synonyms within different sentences or abstracts. One common challenge, especially for biomedical terminology, is the so-called word-within-a-word problem. In particular, many gene, protein and metabolite names contain multiple words that also contain within them other gene, protein or metabolite names. For instance the protein name "Fibroblast growth factor receptor" contains another protein name "Fibroblast growth factor". Likewise the protein name "Glucose-6-phosphate isomerase" contains three metabolite names "Glucose", "phosphate" and "Glucose-6-phosphate". To avoid name mis-identification PolySearch uses the longest-matching phrase approach to screen "overmatched" words and phrases. This involves creating a list of all words and names in the PolySearch thesauruses that contain other PolySearch words and names within them. This list is then used to mask words and/or phrases from the sentences that are longer than the intended (i.e. query/thesaurus) word or phrase. Because of this feature PolySearch can correctly identify the protein "Glucose-6-phosphate isomerase" instead of identifying the metabolite "Glucose-6-phosphate".

PolySearch uses a number of other ad hoc rules to avoid mismatches and to improve its sensitivity and specificity. For instance all names (gene, metabolite, disease or drug) that are less than 3 characters long are excluded from its thesaurus lists. Likewise for any word less than 6 characters long, PolySearch uses a case-sensitive exact match system in order to better identify possible abbreviations. PolySearch also makes intelligent use of dashes and hyphens to help differentiate names within names and avoid the problems arising from different "-" usage.

4.2 PolySearch Sentence Scoring, Ranking and Integration

A central premise to PolySearch is the assumption that the greater the frequency with which an X and Y association occurs within a collection of abstracts or databases, the more significant the association is likely to be. For instance, if COX2 is mentioned as being associated with colon cancer 510 times in PubMed but thioredoxin is associated with colon cancer only once, then one is more likely to have more confidence in the COX2-colon cancer association. This is similar to the concept of ranking papers in terms of their significance by using citation data (as done by ISI) or ranking web pages by their link frequencies (as done by Google [31]). However, citation frequency or link frequency alone is not always the best way to rate a paper or a website for its relevancy. Therefore, in addition to counting the frequency of apparent associations, PolySearch also employs a text ranking scheme to score the most relevant sentences and abstracts that associate both query and words with each other. This is done by assigning a relevancy score to each abstract or pseudo-abstract. As mentioned earlier, each abstract is first divided into individual sentences. Using these individual sentences, PolySearch tries to find query words, association words and thesaurus words in order to classify what we call R1, R2, R3, and R4 sentences (R stands for relevancy). An R4 sentence is a sentence that contains just one of the thesaurus words. The purpose of counting R4 sentence is to provide a measure of the frequency of appearances for the thesaurus words in abstracts mentioning the query word. An R3 sentence is a sentence that has one of the thesaurus words as well as the query word. Compared to an R4 sentence, an R3 sentence is obviously a stronger evidence for association between the query and the thesaurus words. However, R3 sentences alone can still lead to inaccurate associations and R3 sentences do not tell us the type of association. For example, among protein-protein interactions, there can be many types of association words such as "binding", "interacting", "inhibiting", "catalyzing", etc. This is where R2 and R1 sentences come in. An R2 sentence is a sentence that has one of the thesaurus words, one of the query words, as well as one of the association words. An R1 sentence is the same as an R2 sentence but in addition, an R1 sentence has to pass PolySearch's pattern recognition criteria. This kind of pattern recognition has been widely used in other text mining systems (such as ALIBABA) to extract protein-protein interactions [12, 22, 32]. PolySearch's pattern recognition system is rule based. It attempts to capture three main types of patterns: 1) "Query Word-Association Word-Thesaurus Word" (or "Thesaurus Word-Association Word-Query Word"), e.g. A phosphorylates B. 2) "Association Word-Query Word-Thesaurus Word" (or "Association Word-Thesaurus Word-Query Word"), e.g. Interaction of A and B. 3) "Query Word-Thesaurus Word-Association Word" (or "Thesaurus Word-Query Word-Association Word"), e.g. A B complex. It has been reported that for active relations between proteins in the literature, 90% are expressed syntactically as "protein-verb-protein" (i.e. a subset of the "Query Word-Association Word-Thesaurus Word" pattern) [33]. Some of the more important rules in PolySearch's pattern recognition system are as follows:

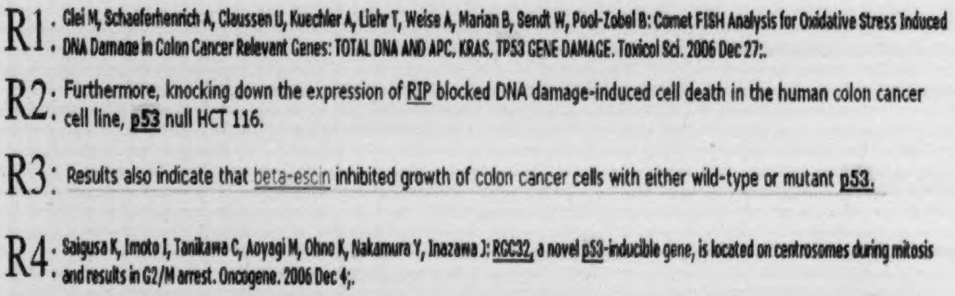
- For the main pattern "Query Word-Association Word-Thesaurus Word", PolySearch searches for compact patterns first. If a compact pattern cannot be found, then PolySearch searches for general patterns. If a general pattern cannot be found, then PolySearch searches for relaxed patterns.
- Compact patterns:
    - The query word and the association word must be within 5 words (tokens) of each other.
    - A "Query Word-Association Word-Thesaurus Word" pattern must be established (i.e. all three types of words are present) within 10 words (tokens) of the query word.
    - A stop word such as "that", "which", "whereas" or "no" cannot be in a "Query Word-Association Word-Thesaurus Word" pattern.
    - Once a "Query Word-Association Word-Thesaurus Word" pattern is established, any thesaurus words that come after that phrase can also meet the pattern recognition criteria.
    - Once a "Query Word-Association Word-Thesaurus Word" pattern is established, if another association word or stop word is seen, the pattern resets.
- General patterns:
    - All relevant words must be within 40 words (tokens) of each other.

    - A “Query Word-Association Word-Thesaurus Word” pattern must be established (i.e. all three types of words are present) within 15 words (tokens) of the query word.
    - A stop word such as “that”, “which”, “whereas” or “no” cannot be in a “Query Word-Association Word-Thesaurus Word” pattern.
    - Once a “Query Word-Association Word-Thesaurus Word” pattern is established, any thesaurus words that come after that phrase can also meet the pattern recognition criteria.
    - Once a “Query Word-Association Word-Thesaurus Word” pattern is established, if another association word or stop word is seen, the pattern resets.
- Relaxed patterns:
    - All relevant words must be within 45 words (tokens) of each other.
    - The query word and the association word must be within 30 words (tokens) of each other.
    - A “Query Word-Association Word-Thesaurus Word” pattern must be established (i.e. all three types of words are present) within 40 words (tokens) of the query word.
    - Once a “Query Word-Association Word-Thesaurus Word” pattern is established, any thesaurus words that come after that phrase can also meet the pattern recognition criteria.

    - Once a “Query Word-Association Word-Thesaurus Word” pattern is established, if another association word is seen, the pattern resets.

- For the “Association Word-Query Word-Thesaurus Word” pattern (mainly for Gene/Protein searches), the association word must have a suffix of -ate, -fer, -ment, -ing, -ion, -lex, -es, or -ions. In addition, all three words must be within 10 words (tokens) of each other.
- For the “Query Word-Thesaurus Word-Association Word” pattern (mainly for Gene/Protein searches), the association word must be one of “complex”, “complexes”, “inhibitor”, “inhibitors”, “interaction”, or “interactions”. In addition, all three words must be within 8 words (tokens) of each other.

The purpose of R1 sentences is to try to capture direct evidence of association between the query word and the thesaurus word, as an R1 sentence provides even stronger evidence of association than an R2 sentence. R1 sentences help ensure that what PolySearch found is correct and R2 sentences ensure that the R1 sentences did not miss anything important. R1 and R2 sentences complement each other to improve the performance of PolySearch. Figure 4.1 shows an example of each type of R (relevancy) sentence where the query is colon cancer.



R1: Glei M, Schaeferhenrich A, Claussen U, Kuechler A, Liehr T, Weise A, Marian B, Sendt W, Pool-Zobel B: Comet FISH Analysis for Oxidative Stress Induced DNA Damage in Colon Cancer Relevant Genes: TOTAL DNA AND APC, KRAS, TP53 GENE DAMAGE. Toxicol Sci. 2006 Dec 27;.

R2: Furthermore, knocking down the expression of RIP blocked DNA damage-induced cell death in the human colon cancer cell line, p53 null HCT 116.

R3: Results also indicate that beta-escin inhibited growth of colon cancer cells with either wild-type or mutant p53.

R4: Saigusa K, Imoto I, Tanikawa C, Aoyagi M, Ohno K, Nakamura Y, Inazawa J: RGC32, a novel p53-inducible gene, is located on centrosomes during mitosis and results in G2/M arrest. Oncogene. 2006 Dec 4;.

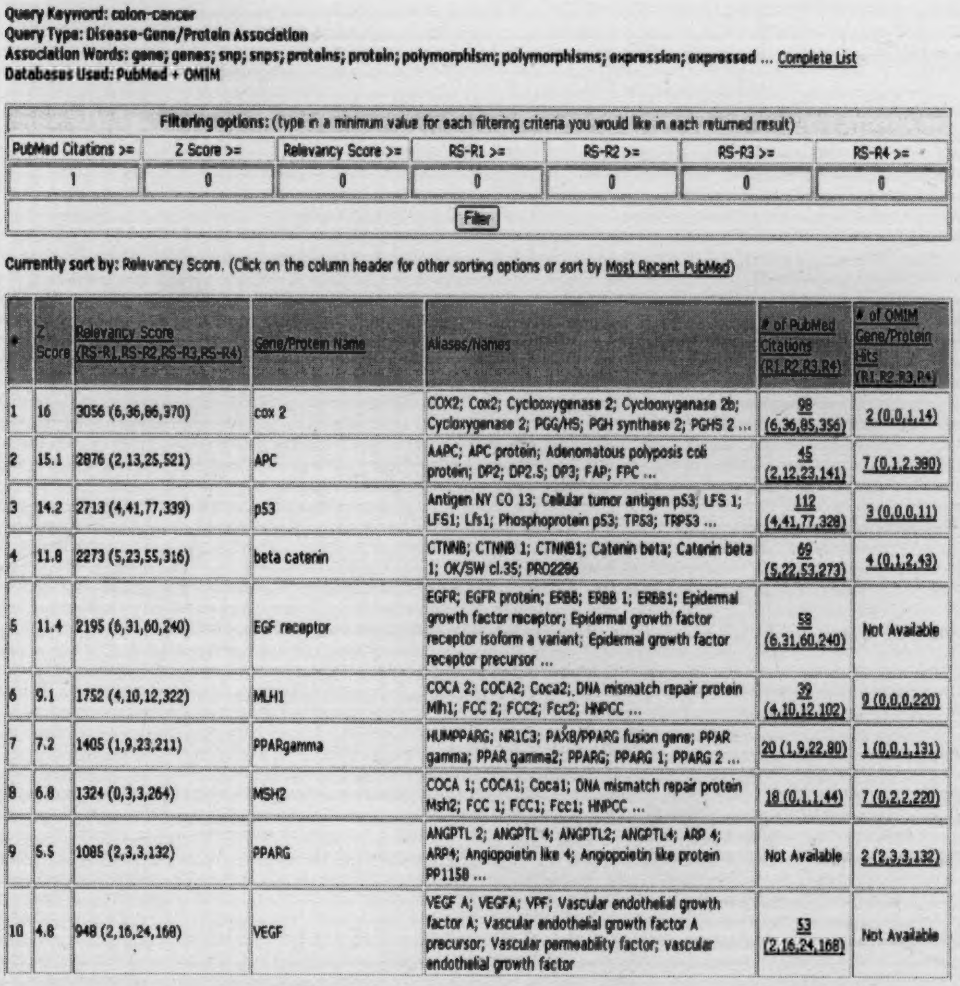
Figure 4.1: An example of each of the R sentence. The query is colon cancer, the association words are coloured in fuchsia, and p53 is the protein of interest.

The total relevancy score is calculated according to the following scoring scheme. For PubMed abstracts, an R4 sentence (weakly relevant) is worth 1 point, an R3 sentence is worth 5 points, an R2 sentence is worth 25 points, and an R1 sentence (strongly relevant) is worth 50 points. For texts from other databases, an R4 sentence is worth 5 points, an R3 sentence is worth 25 points, an R2 sentence is worth 50 points, and an R1 sentence is worth 100 points. The total number of R1, R2, R3, R4 sentences found in both PubMed and each of the other databases are tabulated and a final relevancy score is calculated using the above scoring scheme. Collectively, we call the R1, R2, R3, and R4 sentence counts the PolySearch Relevancy Index (PRI). In the following sections, we will discuss how the PRI can be used like a visual cue as well as a scoring cut-off for indicating and extracting relevant associations.

If the query and/or thesaurus words exist in multiple sentences within the abstract or paragraph the sentence relevancy scores are added together. In this way, an abstract that repeats a given query/thesaurus word association in multiple sentences is given a higher score than an abstract that lists this association only once. Each abstract or paragraph that PolySearch identifies is assigned with this relevancy score. Typically, PolySearch identifies hundreds of different query word associations from its abstract searches. Therefore, if a given gene/drug/disease/metabolite association is mentioned in fifty abstracts then the total relevancy score is the sum of the individual relevancy scores of each of the fifty abstracts.

4.3 PolySearch Results

When a PolySearch query is completed, the results are displayed in an HTML table (see Figure 4.2).



Query Keyword: colon-cancer
Query Type: Disease-Gene/Protein Association
Association Words: gene; genes; snp; snps; proteins; protein; polymorphism; polymorphisms; expression; expressed ... Complete List
Databases Used: PubMed + OMIM

Filtering options: (type in a minimum value for each filtering criteria you would like in each returned result)

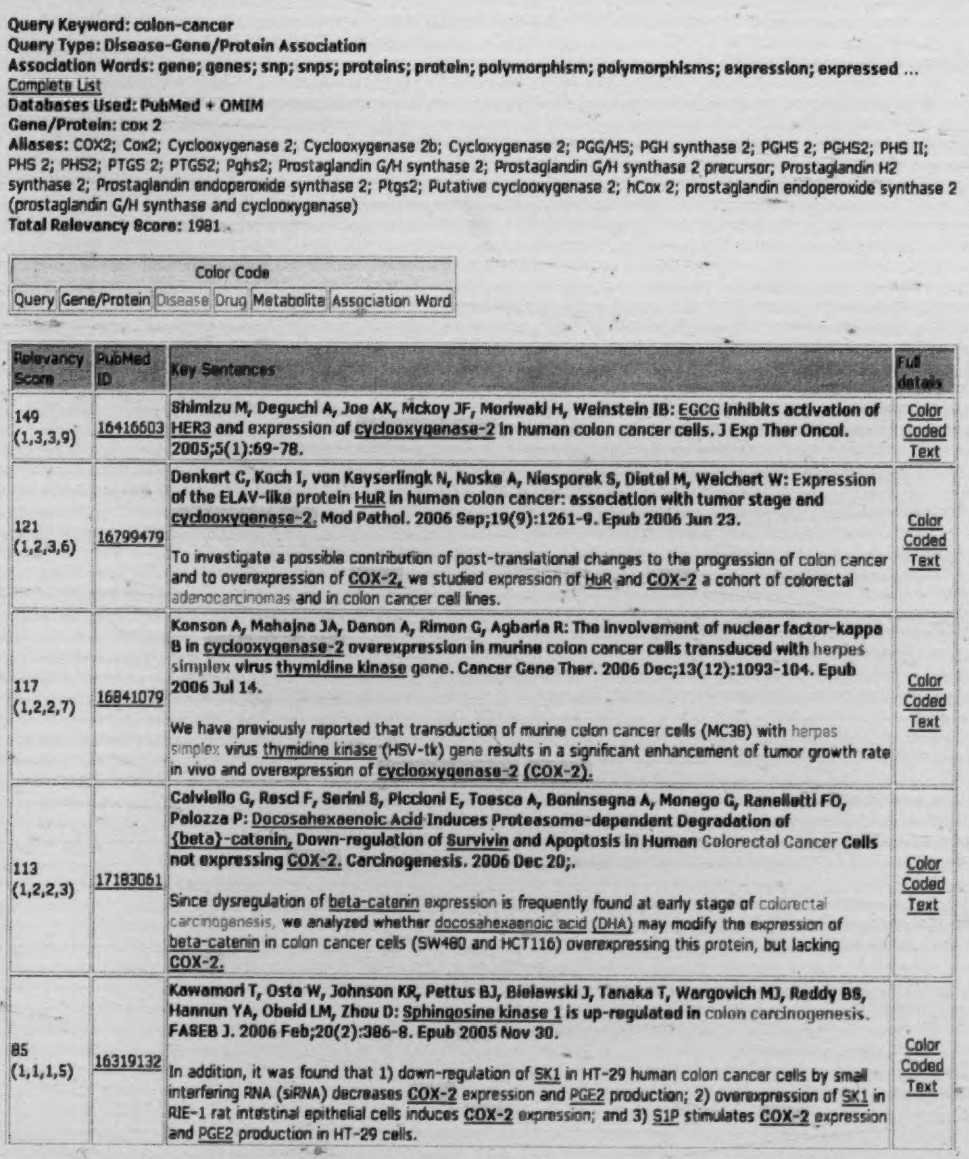
PubMed Citations >=	Z Score >=	Relevancy Score >=	RS-R1 >=	RS-R2 >=	RS-R3 >=	RS-R4 >=
1	0	0	0	0	0	0

Currently sort by: Relevancy Score. (Click on the column header for other sorting options or sort by Most Recent PubMed)

#	Z Score	Relevancy Score (RS-R1,RS-R2,RS-R3,RS-R4)	Gene/Protein Name	Aliases/Names	# of PubMed Citations (R1,R2,R3,R4)	# of OMIM Gene/Protein Hits (R1,R2,R3,R4)
1	16	3056 (6,36,86,370)	cox 2	COX2; Cox2; Cyclooxygenase 2; Cyclooxygenase 2b; Cycloxygenase 2; PGG/HS; PGH synthase 2; PGHS 2 ...	98 (6,36,85,356)	2 (0,0,1,14)
2	15.1	2876 (2,13,25,521)	APC	AAPC; APC protein; Adenomatous polyposis coli protein; DP2; DP2.5; DP3; FAP; FPC ...	45 (2,12,23,141)	7 (0,1,2,380)
3	14.2	2713 (4,41,77,339)	p53	Antigen NY CO 13; Cellular tumor antigen p53; LFS 1; LFS1; Lfs1; Phosphoprotein p53; TP53; TRP53 ...	112 (4,41,77,328)	3 (0,0,0,11)
4	11.8	2273 (5,23,55,316)	beta catenin	CTNNB; CTNNB 1; CTNNB1; Catenin beta; Catenin beta 1; OK/SW cl.35; PRO2286	69 (5,22,53,273)	4 (0,1,2,43)
5	11.4	2195 (6,31,60,240)	EGF receptor	EGFR; EGFR protein; ERBB; ERBB 1; ERBB1; Epidermal growth factor receptor; Epidermal growth factor receptor isoform a variant; Epidermal growth factor receptor precursor ...	58 (6,31,60,240)	Not Available
6	9.1	1752 (4,10,12,322)	MLH1	COCA 2; COCA2; Coca2; DNA mismatch repair protein Mlh1; FCC 2; FCC2; Fcc2; HNPCC ...	39 (4,10,12,102)	9 (0,0,0,220)
7	7.2	1405 (1,9,23,211)	PPARgamma	HUMPPARG; NR1C3; PAX8/PPARG fusion gene; PPAR gamma; PPAR gamma2; PPARG; PPARG 1; PPARG 2 ...	20 (1,9,22,80)	1 (0,0,1,131)
8	6.8	1324 (0,3,3,264)	MSH2	COCA 1; COCA1; Coca1; DNA mismatch repair protein Msh2; FCC 1; FCC1; Fcc1; HNPCC ...	18 (0,1,1,44)	7 (0,2,2,220)
9	5.5	1085 (2,3,3,132)	PPARG	ANGPTL 2; ANGPTL 4; ANGPTL2; ANGPTL4; ARP 4; ARP4; Angiopoietin like 4; Angiopoietin like protein PP1158 ...	Not Available	2 (2,3,3,132)
10	4.8	948 (2,16,24,168)	VEGF	VEGF A; VEGFA; VPF; Vascular endothelial growth factor A; Vascular endothelial growth factor A precursor; Vascular permeability factor; vascular endothelial growth factor	53 (2,16,24,168)	Not Available

Figure 4.2: An example of the output for PolySearch's main results display.

As seen from this figure the top of the table typically summarizes the query word(s), the type of search, the association words used, and the databases used to construct the search. Below the summary is a small table that allows users to filter the results by entering minimum number of PubMed citations, a Z score (a normalized score calculated from the average relevancy score of all hits the query returned), the relevancy score, total R1 count (RS-R1), total R2 count (RS-R2), total R3 count (RS-R3) and total R4 count (RS-R4). Below the filtering options table is the results table (Figure 4.2). The first column in this table indicates the number of hits, while the second column lists each hit's Z score. The results are ranked according to their total relevancy score (by default) which is shown in the third column along with the PolySearch Relevancy Index for total R1, R2, R3, and R4 counts. The fourth column displays the matching thesaurus words while the fifth column displays the corresponding thesaurus synonyms or aliases. The sixth column displays the number of PubMed citations and the PRI for the R1, R2, R3, and R4 counts from PubMed citations. Matches to additional databases (OMIM, HMDB, DrugBank, etc.) are displayed in the remaining columns. Figure 4.2 shows a PolySearch result where both PubMed and OMIM were searched. The results table may be re-sorted by clicking on the column headings or by clicking on the link to sort by the date of the most recent PubMed citation. Clicking the links under the PubMed column (or the other database columns) generates a second HTML table that displays the key sentences found in each database abstract or pseudo abstract along with hyperlinks to the full database record. Data are also provided for the associated PubMed IDs or database accession numbers. For example if 72 PubMed abstracts were identified for a particular drug-gene association, each of the most relevant sentences from every one of the identified abstracts would be displayed in this table. As seen in Figure 4.3, the extracted sentences are colour-coded to facilitate rapid visual scanning. Words marked in red correspond to the query word(s), blue to human genes, green to diseases, brown to drugs, magenta to metabolites and fuchsia to association words (dark yellow is reserved for the other word types such as tissue, organ, subcellular localization and user provided text words).



Query Keyword: colon-cancer
Query Type: Disease-Gene/Protein Association
Association Words: gene; genes; snp; snps; proteins; protein; polymorphism; polymorphisms; expression; expressed ...
Complete List
Databases Used: PubMed + OMIM
Gene/Protein: cox 2
Aliases: COX2; Cox2; Cyclooxygenase 2; Cyclooxygenase 2b; Cycloxygenase 2; PGG/HS; PGH synthase 2; PGHS 2; PGHS2; PHS II; PHS 2; PHS2; PTGS 2; PTGS2; Pghs2; Prostaglandin G/H synthase 2; Prostaglandin G/H synthase 2 precursor; Prostaglandin H2 synthase 2; Prostaglandin endoperoxide synthase 2; Ptgs2; Putative cyclooxygenase 2; hCox 2; prostaglandin endoperoxide synthase 2 (prostaglandin G/H synthase and cyclooxygenase)
Total Relevancy Score: 1981

Color Code					
Query	Gene/Protein	Disease	Drug	Metabolite	Association Word

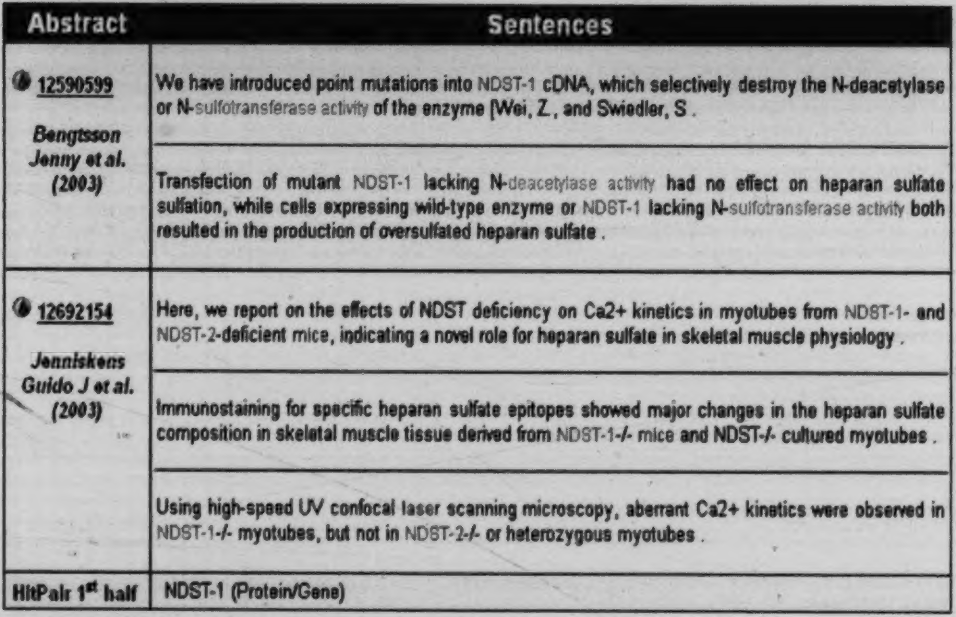
Relevancy Score	PubMed ID	Key Sentences	Full details
149 (1,3,3,9)	16416603	**Shimizu M, Deguchi A, Joe AK, Mckoy JF, Moriwaki H, Weinstein IB: EGCG inhibits activation of HER3 and expression of cyclooxygenase-2 in human colon cancer cells. J Exp Ther Oncol. 2005;5(1):69-78.**	Color Coded Text
121 (1,2,3,6)	16799479	**Denkert C, Koch I, von Keyserlingk N, Noske A, Niesporek S, Dietel M, Weichert W: Expression of the ELAV-like protein HuR in human colon cancer: association with tumor stage and cyclooxygenase-2. Mod Pathol. 2006 Sep;19(9):1261-9. Epub 2006 Jun 23.** To investigate a possible contribution of post-translational changes to the progression of colon cancer and to overexpression of COX-2, we studied expression of HuR and COX-2 a cohort of colorectal adenocarcinomas and in colon cancer cell lines.	Color Coded Text
117 (1,2,2,7)	16841079	**Konson A, Mahajna JA, Danon A, Rimon G, Agbaria R: The involvement of nuclear factor-kappa B in cyclooxygenase-2 overexpression in murine colon cancer cells transduced with herpes simplex virus thymidine kinase gene. Cancer Gene Ther. 2006 Dec;13(12):1093-104. Epub 2006 Jul 14.** We have previously reported that transduction of murine colon cancer cells (MC38) with herpes simplex virus thymidine kinase (HSV-tk) gene results in a significant enhancement of tumor growth rate in vivo and overexpression of cyclooxygenase-2 (COX-2).	Color Coded Text
113 (1,2,2,3)	17183061	**Calviello G, Resci F, Serini S, Piccioni E, Toesca A, Boninsegna A, Monego G, Ranelletti FO, Palozza P: Docosahexaenoic Acid Induces Proteasome-dependent Degradation of {beta}-catenin, Down-regulation of Survivin and Apoptosis in Human Colorectal Cancer Cells not expressing COX-2. Carcinogenesis. 2006 Dec 20;.** Since dysregulation of beta-catenin expression is frequently found at early stage of colorectal carcinogenesis, we analyzed whether docosahexaenoic acid (DHA) may modify the expression of beta-catenin in colon cancer cells (SW480 and HCT116) overexpressing this protein, but lacking COX-2.	Color Coded Text
85 (1,1,1,5)	16319132	**Kawamori T, Osta W, Johnson KR, Pettus BJ, Bielawski J, Tanaka T, Wargovich MJ, Reddy BS, Hannun YA, Obeid LM, Zhou D: Sphingosine kinase 1 is up-regulated in colon carcinogenesis. FASEB J. 2006 Feb;20(2):386-8. Epub 2005 Nov 30.** In addition, it was found that 1) down-regulation of SK1 in HT-29 human colon cancer cells by small interfering RNA (siRNA) decreases COX-2 expression and PGE2 production; 2) overexpression of SK1 in RIE-1 rat intestinal epithelial cells induces COX-2 expression; and 3) S1P stimulates COX-2 expression and PGE2 production in HT-29 cells.	Color Coded Text

Figure 4.3: An example of the key sentences that are extracted and evaluated from a standard PolySearch run.

If a query word happens to be a gene, drug, disease or metabolite, the red colour of the query word takes precedence. Words highlighted with a light yellow background are the current thesaurus words that the user is viewing. This highlighting is used to facilitate rapid visual cueing of the association between the query word and the thesaurus word. The same colour coding scheme is used in PolySearch's fully annotated abstract view.

Perhaps the most useful feature displayed in PolySearch's results page is the PolySearch Relevancy Index which displays the R1, R2, R3, R4 sentence counts. This scoring scheme provides immediate visual cues regarding the quality of association and biological significance. It is the type of information that users most often want but which is not available for most other text mining systems. As a rule of thumb, any association that has an R2 score of at least 1 (meaning one R2 sentence was found which mentions the query word, a association word and a thesaurus word) is worth further investigation. As the R1 and R2 scores get higher, it is more likely that the association found by PolySearch has biological or biomedical significance. In PolySearch's overview page (Figure 4.2), the PRI scoring display serves as a simple cue for the quality of association and biological significance without requiring users to look at the key sentences. In the key sentences page (Figure 4.3), the PRI scoring display provides information about how much more information the abstract has to offer. The key sentences page only displays a single key sentence if any of the R1, R2, or R3 sentences is available (ranked in that order) or two key sentences, with one sentence mentioning the query word and one sentence mentioning the thesaurus word. In other words, if one abstract has more than one R1 or R2 sentence, then it may worth further reading.

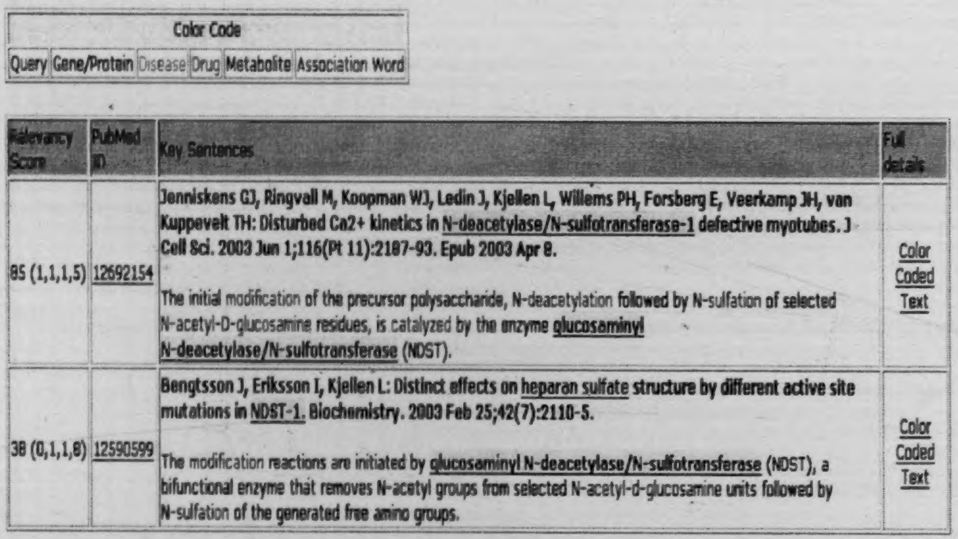
As a demonstration of the effectiveness of using this kind of relevancy scoring, here is a brief sentence comparison between EBIMed and PolySearch with the same query, same gene of interest and same abstracts. The query is "N-acetyl-D-glucosamine" and the gene of interest is "NDST-1" plus its synonyms. Figure 4.4 shows what EBIMed found.



Abstract	Sentences
12590599 *Bengtsson Jenny et al. (2003)*	We have introduced point mutations into NDST-1 cDNA, which selectively destroy the N-deacetylase or N-sulfotransferase activity of the enzyme [Wei, Z., and Swiedler, S.
	Transfection of mutant NDST-1 lacking N-deacetylase activity had no effect on heparan sulfate sulfation, while cells expressing wild-type enzyme or NDST-1 lacking N-sulfotransferase activity both resulted in the production of oversulfated heparan sulfate.
12692154 *Jenniskens Guido J et al. (2003)*	Here, we report on the effects of NDST deficiency on Ca2+ kinetics in myotubes from NDST-1- and NDST-2-deficient mice, indicating a novel role for heparan sulfate in skeletal muscle physiology.
	Immunostaining for specific heparan sulfate epitopes showed major changes in the heparan sulfate composition in skeletal muscle tissue derived from NDST-1-/- mice and NDST-/- cultured myotubes.
	Using high-speed UV confocal laser scanning microscopy, aberrant Ca2+ kinetics were observed in NDST-1-/- myotubes, but not in NDST-2-/- or heterozygous myotubes.
HitPair 1st half	NDST-1 (Protein/Gene)

Figure 4.4: The key sentences EBIMed found for the query N-acetyl-D-glucosamine and gene of interest is NDST-1.

The EBIMed score for the results shown in Figure 4.4 is 5, which is quite low. Looking at the key sentences, the word N-acetyl-D-glucosamine cannot even be found so it is likely the reader is would think that there is no association between N-acetyl-D-glucosamine and NDST-1. In contrast, Figure 4.5 shows what PolySearch found using the exact same two abstracts.



Color Code					
Query	Gene/Protein	Disease	Drug	Metabolite	Association Word

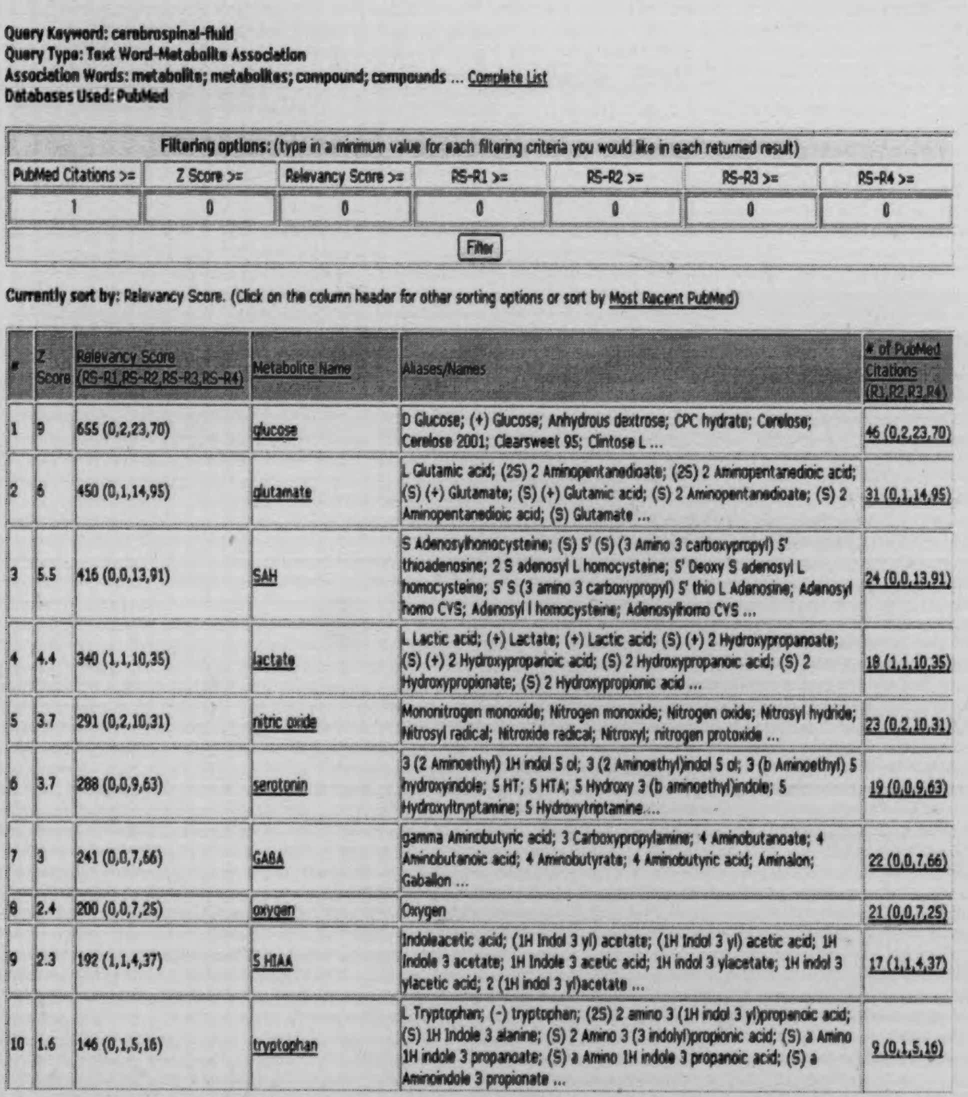
Relevancy Score	PubMed ID	Key Sentences	Full details
85 (1,1,1,5)	12692154	**Jenniskens GJ, Ringvall M, Koopman WJ, Ledin J, Kjellen L, Willems PH, Forsberg E, Veerkamp JH, van Kuppevelt TH: Disturbed Ca2+ kinetics in N-deacetylase/N-sulfotransferase-1 defective myotubes. J Cell Sci. 2003 Jun 1;116(Pt 11):2187-93. Epub 2003 Apr 8.** The initial modification of the precursor polysaccharide, N-deacetylation followed by N-sulfation of selected N-acetyl-D-glucosamine residues, is catalyzed by the enzyme glucosaminyl N-deacetylase/N-sulfotransferase (NDST).	Color Coded Text
38 (0,1,1,8)	12590599	**Bengtsson J, Eriksson I, Kjellen L: Distinct effects on heparan sulfate structure by different active site mutations in NDST-1. Biochemistry. 2003 Feb 25;42(7):2110-5.** The modification reactions are initiated by glucosaminyl N-deacetylase/N-sulfotransferase (NDST), a bifunctional enzyme that removes N-acetyl groups from selected N-acetyl-d-glucosamine units followed by N-sulfation of the generated free amino groups.	Color Coded Text

Figure 4.5: The key sentences PolySearch found for the query N-acetyl-D-glucosamine and gene of interest is NDST-1.

The PolySearch score for the result shown in Figure 4.5 is 123 with a PRI of (1,2,2,13) (R1 = 1, R2 = 2, R3 = 3, R4 = 13). By reading the key sentences that PolySearch found, the reader is more likely to ascertain that there is an association between N-acetyl-D-glucosamine and NDST-1 or at the very least, that this association is worth further investigation. As this example has shown, identifying the query words and the association words inside sentences is key to extracting meaningful information.

4.4 Improve Association Word Selections/Relevance Feedback

One of the unique features in PolySearch is its flexible association-word ranking (relevance feedback) system. A list of association words can be used in a PolySearch query in order to determine how the results returned by PolySearch are to be ranked. Different types of searches require different sets of association words and a default set of association words is embedded in PolySearch to assist users for each of the different types of searches that PolySearch supports. Users can also edit the list of association words to suit their needs and further refine the search. Choosing the proper association words is essential in PolySearch's ranking algorithm. Here we will use a relatively uncommon search, "Given cerebrospinal fluid or CSF Find Metabolites", to demonstrate how to improve the association word selection to refine a given search. To begin, one can simply guess some common some association words on their own such as "metabolite, metabolites, compound, compounds" or alternately one can choose to use no association words. Figure 4.6 shows the results that PolySearch returned using cerebrospinal fluid as the query and "metabolite, metabolites, compound, compounds" as the association words.



Query Keyword: cerebrospinal-fluid
Query Type: Text Word-Metabolite Association
Association Words: metabolite; metabolites; compound; compounds ... Complete List
Databases Used: PubMed

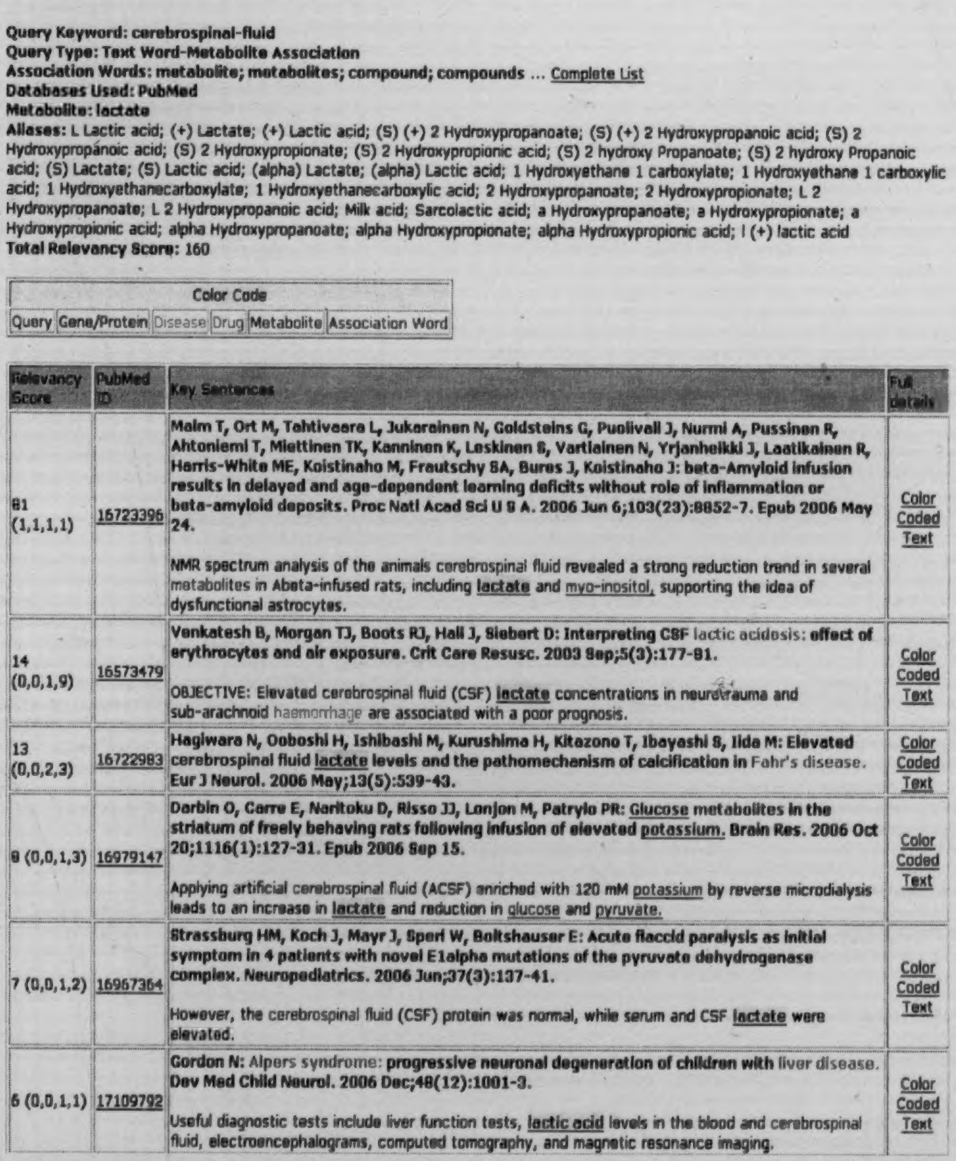
Filtering options: (type in a minimum value for each filtering criteria you would like in each returned result)						
PubMed Citations >=	Z Score >=	Relevancy Score >=	RS-R1 >=	RS-R2 >=	RS-R3 >=	RS-R4 >=
1	0	0	0	0	0	0

Currently sort by: Relevancy Score. (Click on the column header for other sorting options or sort by Most Recent PubMed)

#	Z Score	Relevancy Score (RS-R1,RS-R2,RS-R3,RS-R4)	Metabolite Name	Aliases/Names	# of PubMed Citations (R1,R2,R3,R4)
1	9	655 (0,2,23,70)	glucose	D Glucose; (+) Glucose; Anhydrous dextrose; CPC hydrate; Cerelose; Cerelose 2001; Clearsweet 95; Clintose L ...	46 (0,2,23,70)
2	6	450 (0,1,14,95)	glutamate	L Glutamic acid; (2S) 2 Aminopentanedioate; (2S) 2 Aminopentanedioic acid; (S) (+) Glutamate; (S) (+) Glutamic acid; (S) 2 Aminopentanedioate; (S) 2 Aminopentanedioic acid; (S) Glutamate ...	31 (0,1,14,95)
3	5.5	416 (0,0,13,91)	SAH	S Adenosylhomocysteine; (S) 5' (S) (3 Amino 3 carboxypropyl) 5' thioadenosine; 2 S adenosyl L homocysteine; 5' Deoxy S adenosyl L homocysteine; 5' S (3 amino 3 carboxypropyl) 5' thio L Adenosine; Adenosyl homo CYS; Adenosyl l homocysteine; Adenosylhomo CYS ...	24 (0,0,13,91)
4	4.4	340 (1,1,10,35)	lactate	L Lactic acid; (+) Lactate; (+) Lactic acid; (S) (+) 2 Hydroxypropanoate; (S) (+) 2 Hydroxypropanoic acid; (S) 2 Hydroxypropanoic acid; (S) 2 Hydroxypropionate; (S) 2 Hydroxypropionic acid ...	18 (1,1,10,35)
5	3.7	291 (0,2,10,31)	nitric oxide	Mononitrogen monoxide; Nitrogen monoxide; Nitrogen oxide; Nitrosyl hydride; Nitrosyl radical; Nitroxide radical; Nitroxyl; nitrogen protoxide ...	23 (0,2,10,31)
6	3.7	288 (0,0,9,63)	serotonin	3 (2 Aminoethyl) 1H indol 5 ol; 3 (2 Aminoethyl)indol 5 ol; 3 (b Aminoethyl) 5 hydroxyindole; 5 HT; 5 HTA; 5 Hydroxy 3 (b aminoethyl)indole; 5 Hydroxyltryptamine; 5 Hydroxytriptamine ...	19 (0,0,9,63)
7	3	241 (0,0,7,66)	GABA	gamma Aminobutyric acid; 3 Carboxypropylamine; 4 Aminobutanoate; 4 Aminobutanoic acid; 4 Aminobutyrate; 4 Aminobutyric acid; Aminalon; Gaballon ...	22 (0,0,7,66)
8	2.4	200 (0,0,7,25)	oxygen	Oxygen	21 (0,0,7,25)
9	2.3	192 (1,1,4,37)	5 HIAA	Indoleacetic acid; (1H Indol 3 yl) acetate; (1H Indol 3 yl) acetic acid; 1H Indole 3 acetate; 1H Indole 3 acetic acid; 1H indol 3 ylacetate; 1H indol 3 ylacetic acid; 2 (1H indol 3 yl)acetate ...	17 (1,1,4,37)
10	1.6	146 (0,1,5,16)	tryptophan	L Tryptophan; (-) tryptophan; (2S) 2 amino 3 (1H indol 3 yl)propenoic acid; (S) 1H Indole 3 alanine; (S) 2 Amino 3 (3 indolyl)propionic acid; (S) a Amino 1H indole 3 propanoate; (S) a Amino 1H indole 3 propanoic acid; (S) a Aminoindole 3 propionate ...	9 (0,1,5,16)

Figure 4.6: Results from a "Given Text Word Find Metabolites" query where the query is cerebrospinal fluid, the association words are metabolite, metabolites, compound and compounds, and the maximum number of abstracts is set to 2000.

Examining the results in Figure 4.6, we can see that in the rightmost column that there are number of hits that has R1, R2 scores ranging from 0 to 2 and R3 scores ranging from 4 to 23. The goal here is try to make an R3 sentence becomes an R1 or an R2 sentence. So, one can try to examine the key sentences of a hit that has high R3 score and low R1 or R2 score such as the one shown in Figure 4.7 (the key sentences for lactate).



Query Keyword: cerebrospinal-fluid
Query Type: Text Word-Metabolite Association
Association Words: metabolite; metabolites; compound; compounds ... Complete List
Databases Used: PubMed
Metabolite: lactate
Aliases: L Lactic acid; (+) Lactate; (+) Lactic acid; (S) (+) 2 Hydroxypropanoate; (S) (+) 2 Hydroxypropanoic acid; (S) 2 Hydroxypropanoic acid; (S) 2 Hydroxypropionate; (S) 2 Hydroxypropionic acid; (S) 2 hydroxy Propanoate; (S) 2 hydroxy Propanoic acid; (S) Lactate; (S) Lactic acid; (alpha) Lactate; (alpha) Lactic acid; 1 Hydroxyethane 1 carboxylate; 1 Hydroxyethane 1 carboxylic acid; 1 Hydroxyethanecarboxylate; 1 Hydroxyethanecarboxylic acid; 2 Hydroxypropanoate; 2 Hydroxypropionate; L 2 Hydroxypropanoate; L 2 Hydroxypropanoic acid; Milk acid; Sarcolactic acid; a Hydroxypropanoate; a Hydroxypropionate; a Hydroxypropionic acid; alpha Hydroxypropanoate; alpha Hydroxypropionate; alpha Hydroxypropionic acid; l (+) lactic acid
Total Relevancy Score: 160

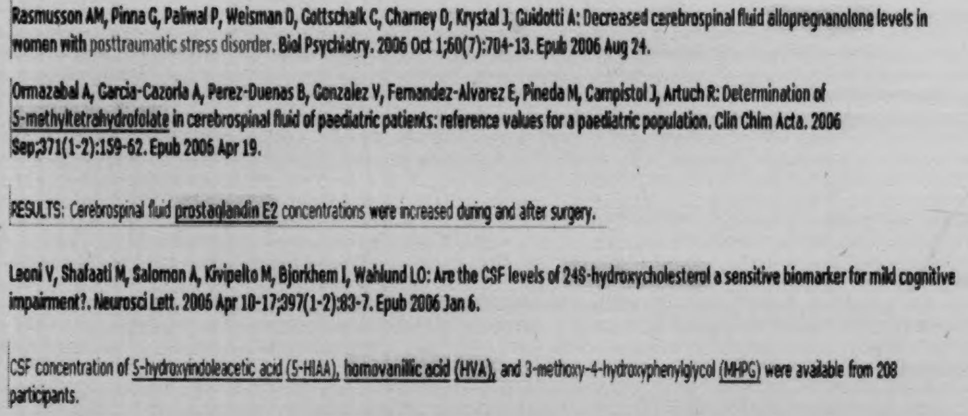
Color Code					
Query	Gene/Protein	Disease	Drug	Metabolite	Association Word

Relevancy Score	PubMed ID	Key Sentences	Full details
81 (1,1,1,1)	16723396	**Malm T, Ort M, Tahtivaara L, Jukarainen N, Goldsteins G, Puolivali J, Nurmi A, Pussinen R, Ahtoniemi T, Miettinen TK, Kanninen K, Leskinen S, Vartiainen N, Yrjanheikki J, Laatikainen R, Harris-White ME, Koistinaho M, Frautschy SA, Bures J, Koistinaho J: beta-Amyloid infusion results in delayed and age-dependent learning deficits without role of inflammation or beta-amyloid deposits. Proc Natl Acad Sci U S A. 2006 Jun 6;103(23):8852-7. Epub 2006 May 24.** NMR spectrum analysis of the animals cerebrospinal fluid revealed a strong reduction trend in several metabolites in Abeta-infused rats, including lactate and myo-inositol, supporting the idea of dysfunctional astrocytes.	Color Coded Text
14 (0,0,1,9)	16573479	**Venkatesh B, Morgan TJ, Boots RJ, Hall J, Siebert D: Interpreting CSF** lactic acidosis: **effect of erythrocytes and air exposure. Crit Care Resusc. 2003 Sep;5(3):177-81.** OBJECTIVE: Elevated cerebrospinal fluid (CSF) lactate concentrations in neurotrauma and sub-arachnoid haemorrhage are associated with a poor prognosis.	Color Coded Text
13 (0,0,2,3)	16722983	**Hagiwara N, Ooboshi H, Ishibashi M, Kurushima H, Kitazono T, Ibayashi S, Iida M: Elevated cerebrospinal fluid lactate levels and the pathomechanism of calcification in** Fahr's disease. **Eur J Neurol. 2006 May;13(5):539-43.**	Color Coded Text
8 (0,0,1,3)	16979147	**Darbin O, Carre E, Naritoku D, Risso JJ, Lanjon M, Patrylo PR: Glucose metabolites in the striatum of freely behaving rats following infusion of elevated potassium. Brain Res. 2006 Oct 20;1116(1):127-31. Epub 2006 Sep 15.** Applying artificial cerebrospinal fluid (ACSF) enriched with 120 mM potassium by reverse microdialysis leads to an increase in lactate and reduction in glucose and pyruvate.	Color Coded Text
7 (0,0,1,2)	16967364	**Strassburg HM, Koch J, Mayr J, Sperl W, Boltshauser E: Acute flaccid paralysis as initial symptom in 4 patients with novel E1alpha mutations of the pyruvate dehydrogenase complex. Neuropediatrics. 2006 Jun;37(3):137-41.** However, the cerebrospinal fluid (CSF) protein was normal, while serum and CSF lactate were elevated.	Color Coded Text
6 (0,0,1,1)	17109792	**Gordon N:** Alpers syndrome: **progressive neuronal degeneration of children with** liver disease. **Dev Med Child Neurol. 2006 Dec;48(12):1001-3.** Useful diagnostic tests include liver function tests, lactic acid levels in the blood and cerebrospinal fluid, electroencephalograms, computed tomography, and magnetic resonance imaging.	Color Coded Text

Figure 4.7: The key sentences for lactate for the "Given Cerebrospinal Fluid Find Metabolites" search.

Looking at the key sentences, one can quickly note that the word "concentrations" could be a good association word (from the key sentence of

PMID 16573479), “levels” could be another good association word (from the key sentence of PMID 16722983) and “elevated” could be another useful association word (from the key sentence of PMID 16967364). Repeating this process for other metabolites and repeating a PolySearch query using a different set of association words will allow the list of association words to start to build up with the more relevant metabolites in cerebrospinal fluid start to move up in PolySearch’s ranking. After 3~4 iterations of this association word selection/refinement, the following association words were found to be effective: “accumulate”, “amount”, “analysis”, “analyse”, “analyze”, “assay”, “component”, “compound”, “concentration”, “contain”, “decline”, “decrease”, “detect”, “determine”, “elevate”, “exceed”, “extract”, “excrete”, “find”, “higher”, “increase”, “identify”, “level”, “localize”, “lower”, “measure”, “metabolite”, “presence”, “purify”, “quantitate”, and “quantify”. Figure 4.8 shows some examples of the results that PolySearch found using this new set of association words.



Rasmusson AM, Pinna G, Paliwal P, Weisman D, Gottschalk C, Charney D, Krystal J, Guidotti A: Decreased cerebrospinal fluid allopregnanolone levels in women with posttraumatic stress disorder. Biol Psychiatry. 2006 Oct 1;60(7):704-13. Epub 2006 Aug 24.

Ormazabal A, Garcia-Cazorla A, Perez-Duenas B, Gonzalez V, Fernandez-Alvarez E, Pineda M, Campistol J, Artuch R: Determination of 5-methyltetrahydrofolate in cerebrospinal fluid of paediatric patients: reference values for a paediatric population. Clin Chim Acta. 2006 Sep;371(1-2):159-62. Epub 2006 Apr 19.

RESULTS: Cerebrospinal fluid prostaglandin E2 concentrations were increased during and after surgery.

Leoni V, Shafaati M, Salomon A, Kivipelto M, Bjorkhem I, Wahlund LO: Are the CSF levels of 24S-hydroxycholesterol a sensitive biomarker for mild cognitive impairment?. Neurosci Lett. 2006 Apr 10-17;397(1-2):83-7. Epub 2006 Jan 6.

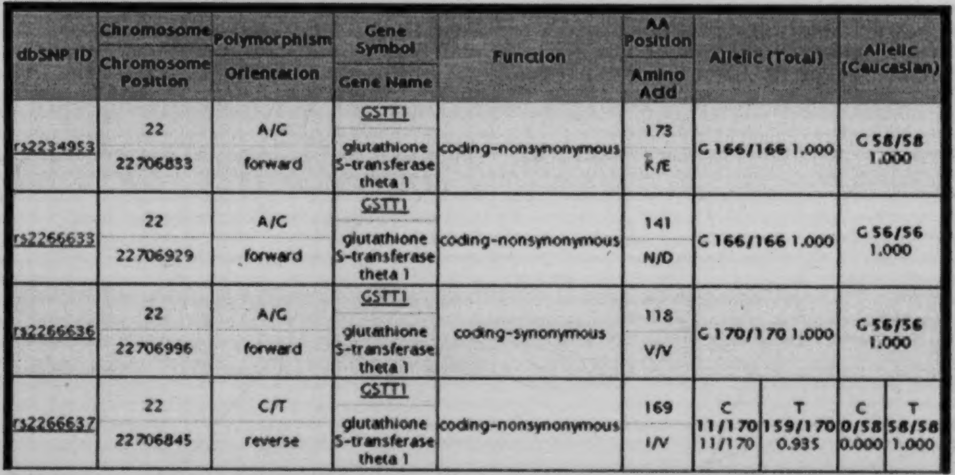
CSF concentration of 5-hydroxyindoleacetic acid (5-HIAA), homovanillic acid (HVA), and 3-methoxy-4-hydroxyphenylglycol (MHPG) were available from 208 participants.

Figure 4.8: Some example R1 or R2 sentences that PolySearch found for the “Given Cerebrospinal Fluid Find Metabolites” search. These examples were found while briefly browsing through the results.

This example demonstrates that PolySearch can be used as a text mining tool to identify optimal association words too. Once an association word list is compiled, one can reuse the association words for similar searches. This flexibility of allowing users to decide what is more relevant along with tools to help create the association word list are examples of features that are unique to PolySearch.

4.5 SNPs

One of the original motivations in developing PolySearch was to facilitate the identification of disease-associated polymorphisms (SNPs) and disease-associated mutations in humans. Discussions with users and SNP researchers provided directions on what SNP or mutation data would be most relevant, what SNP databases would contain the most relevant data and what interface features would be most useful. The results of these discussions led to the design of several data viewing tables which can be seen in Figures 4.9 and 4.10. Figure 4.9 shows the results of a "Given SNP Find Associated Gene" query and the list of SNP features that PolySearch returned. These include: 1) the RS (reference SNP) number, 2) chromosome number, 3) chromosome position, 4) polymorphism (e.g. A to G), 5) strand orientation, 6) gene symbol, 7) gene name, 8) SNP function (synonymous, nonsynonymous, deleterious, etc), 9) amino acid position, 10) amino acid and 11) allele frequencies (total or Caucasian). The SNP data is gathered using web mining techniques similar to those used by BioSpider [34] and the data is derived from the following databases: HapMap [28], Entrez SNP (db SNP) [3], CGAP [29], and HGMD [5]. Users can enter a gene name, gene symbol, gene sequence, or an RS number to retrieve the SNP data and view them in an HTML table (Figure 4.9) or download these data into a MS Excel file for further analysis.

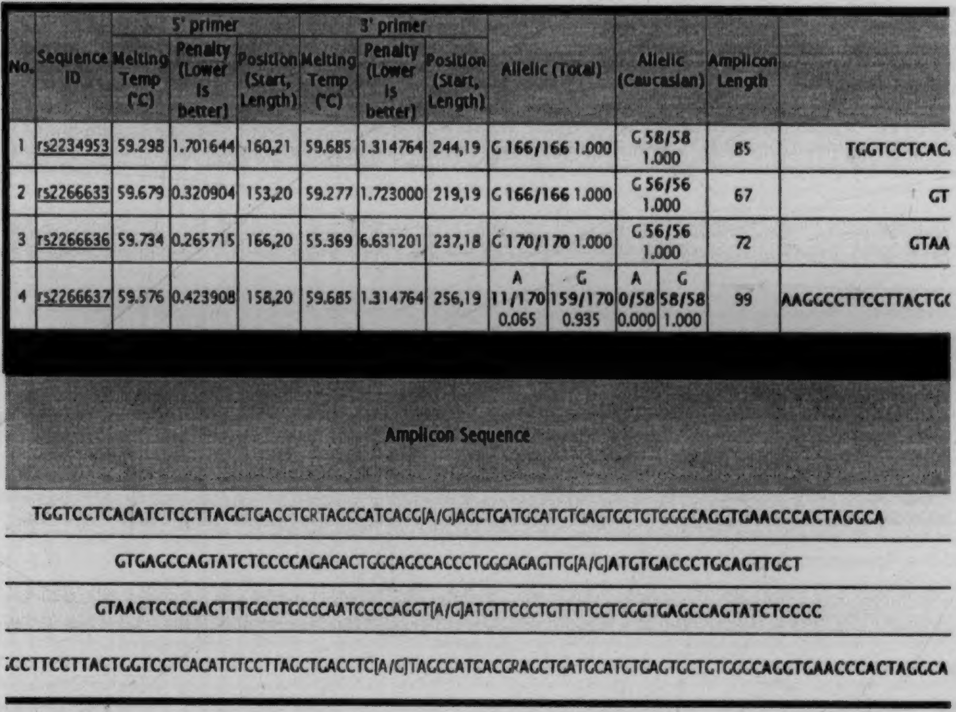


dbSNP ID	Chromosome	Polymorphism	Gene Symbol	Function	AA Position	Allelic (Total)		Allelic (Caucasian)	
	Chromosome Position	Orientation	Gene Name		Amino Acid				
rs2234953	22 22706833	A/G forward	GSTT1 glutathione S-transferase theta 1	coding-nonsynonymous	173 K/E	G 166/166 1.000		G 58/58 1.000	
rs2266633	22 22706929	A/G forward	GSTT1 glutathione S-transferase theta 1	coding-nonsynonymous	141 N/D	G 166/166 1.000		G 56/56 1.000	
rs2266636	22 22706996	A/G forward	GSTT1 glutathione S-transferase theta 1	coding-synonymous	118 V/V	G 170/170 1.000		G 56/56 1.000	
rs2266637	22 22706845	C/T reverse	GSTT1 glutathione S-transferase theta 1	coding-nonsynonymous	169 I/V	C 11/170 11/170	T 159/170 0.935	C 0/58 0.000	T 58/58 1.000

Figure 4.9: The output for a "SNP to Gene" search using rs2234953, rs2266633, rs2266636, and rs2266637 as input. PolySearch collects important information about SNP such as: position, type of polymorphism, gene symbol/name, function and allelic frequency.

4.6 Primer Design

In many cases the identification of a SNP or a mutation of interest requires further sequence analysis in order to design the experiments needed to detect and confirm these SNPs. Specifically, the capacity to design PCR primers for gene cloning or SNP/mutation detection is particularly important. Users can design PCR primers through PolySearch either by entering a SNP ID (i.e. a RS #) or a gene sequence. Figure 4.10 shows an example of how a PCR primer was designed by PolySearch using only the RS #. One feature that is particularly unique to PolySearch's primer design is that it implements proximity SNP checking. This ensures that the 5' primer and the 3' primer do not overlap with any neighbouring SNPs as primers. This proximity SNP checking ensures that the 5' primer and the 3' primer excludes regions that would affect annealing of primers due to embedded SNPs on the target sequence for which the primers bind. Lack of this information and attention may result in poor performance of PCR primers in the amplification of target sequences. The primer3 program (release 1.0) from the Whitehead Institute for Biomedical Research [35] is used to facilitate PolySearch's PCR primer design. PCR primer design can either be fully automated or semi-automated. In the semi-automatic mode users can choose to enter their own amplicon sequence length, primer length, melting temperature and salt concentration. In the automated mode, gene sequences (+/- 200 bp of the query polymorphism site) are identified by PolySearch using the RS #. The "TARGET" tag for the primer3 program is used to ensure that the single nucleotide polymorphism site is in the amplicon sequence and the "EXCLUDED_REGION" tag is used to ensure there are no other polymorphism sites in the 3' and 5' primers. As can be seen in Figure 4.10, PolySearch's output format is very informative. Potential primer sequences are identified in bold, the SNP position and polymorphism is clearly identified in square brackets, and the neighbouring SNPs within the amplicon are also identified by differently coloured letters. The amplicon size and the start and end positions of the target sequence are also indicated.



No.	Sequence ID	5' primer Melting Temp (°C)	5' primer Penalty (Lower is better)	5' primer Position (Start, Length)	3' primer Melting Temp (°C)	3' primer Penalty (Lower is better)	3' primer Position (Start, Length)	Allelic (Total)	Allelic (Caucasian)	Amplicon Length	
1	rs2234953	59.298	1.701644	160,21	59.685	1.314764	244,19	G 166/166 1.000	G 58/58 1.000	85	TGGTCCTCAC
2	rs2266633	59.679	0.320904	153,20	59.277	1.723000	219,19	G 166/166 1.000	G 56/56 1.000	67	GT
3	rs2266636	59.734	0.265715	166,20	55.369	6.631201	237,18	G 170/170 1.000	G 56/56 1.000	72	GTAA
4	rs2266637	59.576	0.423908	158,20	59.685	1.314764	256,19	A 11/170 0.065 G 159/170 0.935	A 0/58 0.000 G 58/58 1.000	99	AAGGCCTTCCTTACTG

Amplicon Sequence
TGGTCCTCACATCTCCTTAGCTGACCTCRTAGCCATCACG[A/G]AGCTGATGCATGTCAGTGCTGTGGCCAGGTGAACCCACTAGGCA
GTGAGCCAGTATCTCCCCAGACACTGGCAGCCACCCTGGCAGAGTTG[A/G]ATGTGACCCTGCAGTTGCT
GTAACTCCCGACTTTGCCTGCCCAATCCCCAGGT[A/G]ATGTTCCCTGTTTTCCTGGGTGAGCCAGTATCTCCCC
CCTTCCTTACTGGTCCTCACATCTCCTTAGCTGACCTC[A/G]TAGCCATCACGRAGCTGATGCATGTCAGTGCTGTGGCCAGGTGAACCCACTAGGCA

Figure 4.10: An illustration of the PCR Primer Design feature in PolySearch.

Chapter 5

5. Evaluation

5.1 Results

A text mining tool is only useful if it gives accurate results and extensive coverage in less time than what could be performed using alternative (i.e. non-computational) or competing computational methods. To evaluate PolySearch's performance, we assessed it using eight different tests or methods. These included 1) a comparison of features and capabilities between PolySearch and other biomedical text mining tools; 2) a comparative evaluation of gene synonym identification; 3) an evaluation of PolySearch's ability to identify protein-protein interactions; 4) an evaluation of PolySearch's ability to identify drug/gene associations; 5) an evaluation of PolySearch's ability to identify metabolite/gene associations; 6) an evaluation of PolySearch's ability to identify disease/gene associations; 7) a comparison of speed and coverage between PolySearch and a senior undergraduate student using literature and computer aid queries for a defined search problem; 8) an assessment of PolySearch's ability to perform an integrated disease to PCR primer design search.

5.1.1 Feature Comparison

In the first assessment, PolySearch was compared to seven other well known biomedical text mining tools, namely Entrez [3], MedMiner [8], MedGene [9], LitMiner [10], ALIBABA [12], IHOP [11] and EBIMed [13]. Comparisons included the types of searches supported, the extent of hyperlinking, the presence of access restrictions, the capacity for text and sentence highlighting, the use of word co-occurrence for scoring, support for keywords/association words or pattern recognition, and the number of database integrations for each tool. As seen in Table 5.1, Entrez offers the most extensive database and search coverage as well as the broadest hyperlinking capabilities. However, Entrez is more of an information retrieval system rather than a text mining system and so it lacks the ranking, scoring and sentence highlighting capabilities of other text mining tools. In contrast, MedMiner provides key sentence highlighting capability and organizes these sentences into twelve general categories, which shorten the time required to gather relevant information from the selected texts. However, MedMiner searches are mainly limited to one-to-one searches (e.g. “one gene” to “one drug” search) and this limits the general utility of MedMiner. Both MedGene and LitMiner provide the capability to perform “given X find all Y” types of searches and both of them provide statistical rankings for the associations they found. Nevertheless, both MedGene and LitMiner lack the ability to perform text and sentence highlighting, making it difficult to verify the associations that MedGene and LitMiner found. ALIBABA, IHOP, EBIMed, and PolySearch all have the ability to rank the associations they found and supply both text and sentence highlighting for quick verification of the associations. ALIBABA treats PubMed abstracts and associations as a graph and provides a graphical interface to display the associations it found. While this approach may be useful for a small number of abstracts, for larger numbers of abstracts, the graph becomes almost unusable due to the over-abundance of information. IHOP, while great for identifying Gene/Protein to Gene/Protein interactions, lacks support for many other search types and this limits IHOP's usefulness in other areas of biomedical research. Unlike IHOP, EBIMed provides a means of an analysis that is independent of the initial keyword query and is more flexible with the types of searchers it allows. However, EBIMed uses a pure word co-occurrence approach to assess associations and so it tends to lack the accuracy of systems that use both keywords and pattern matching (such as ALIBABA, IHOP and PolySearch). PolySearch, being the most recent addition, combines some of the best features from each of the other tools. In addition, PolySearch appears to be unique in terms of the diversity of its search and text ranking possibilities, its ability to perform extensive query synonym expansion using its different thesauruses, its PolySearch Relevancy Index (PRI) scoring display for immediate visual indications on the strength of association, its SNP search functionalities, and its ability to text mine additional databases such as OMIM, SwissProt, DrugBank, HMDB, HPRD and GAD.

	Entrez	**MedMiner**	**MedGene**	**LitMiner**	**Alibaba**	**IHOP**	**EBIMed**	**PolySearch**
Type of Search supported	Literature, Disease, Gene, Structure, Taxonomy, SNP, Compound, etc.	Gene, Drug, Text Word	Gene, Disease	Gene, Disease, Compounds, Tissues/Organs	Gene, Disease, Drug, Tissues/Organs, Cells, Species	Gene	Gene, Cellular Compartment, Biological Process, Molecular Function, Drug, Species	Gene, Disease, Drug, Metabolite, Tissues/Organs, Subcellular Localization, Text Word
Extensive hyperlinking	Most Extensive	Less extensive	Less extensive	Less extensive	Less extensive	More Extensive	More Extensive	More Extensive
Access restrictions	None	None	Registration	None	None	None	None	None
Text and sentence highlighting	No	Yes	No	No	Yes	Yes	Yes	Yes
Co-occurrence scoring scheme	None	None	Abstract level	Abstract level	Sentence level	Sentence level	Sentence level	Sentence level
Use of keywords or association words	None	Predefined keywords	None	None	Predefined keywords	Predefined keywords	None	Predefined & custom association words
Sentence pattern recognition	No	No	No	No	Yes	Yes	No	Yes
Thesaurus query synonym expansion	Yes, limited	Yes, limited	Yes, limited	None	None	Yes, for genes only	None	Yes, extensive
Databases	PubMed, OMIM, Gene, MMDB, Taxonomy, dbSNP, PubChem, etc.	PubMed, GeneCards	PubMed	PubMed	PubMed	PubMed, HPRD, IntAct	PubMed	PubMed, OMIM, Swissprot, DrugBank, HMDB, HPRD, GAD, HapMap, dbSNP, CGAP, HGMD

Table 5.1: Feature comparison of various biomedical text mining tools.

5.1.2 Evaluation of Gene/Protein Synonym Identification

For the second assessment, we tested PolySearch's ability to identify genes and protein names within different sentences or abstracts. To do this, we use the dataset that IHOP used for evaluating their gene synonym identification for human genes [36]. The dataset contains 181 sentences from various PubMed abstracts with an average of about 2-3 gene names per sentence (the names include symbols, standard names, abbreviations and synonyms). We manually identified the correct gene and protein names from the dataset and used this collection as our gold standard to compare to PolySearch's gene synonym identification for the dataset. Table 5.2 shows PolySearch's precision, recall and f-measure in this evaluation as compared to IHOP.

	IHOP	PolySearch
Precision (%)	87.1	90.1
Recall (%)	81.8	85.3
F-measure (%)	84.4	87.6

Table 5.2: Precision, recall and f-measure on gene synonym identification for PolySearch and IHOP.

As Table 5.2 shows, PolySearch's performance on gene/protein identification is comparable to that of IHOP, which is generally regarded as one of the better available tools for identifying genes and gene-gene associations [11]. This assessment shows that PolySearch is capable of identifying gene/protein terms with a high level of accuracy, the first important step before trying to extract associations.

5.1.3 Evaluation of Protein-Protein Interactions

In the third assessment, we looked at the performance of PolySearch on protein-protein interaction extraction. This assessment was divided into two parts. First, we evaluated the efficacy of PolySearch's association words and pattern recognition system by using a protein-protein interaction corpus. The protein-protein interaction corpus consisted of the data used in developing PolySearch's rule based pattern recognition system. For the second part, we compared the performance of PolySearch on protein-protein interaction extraction with actual abstracts against two other text mining systems (EBIMed and IHOP), and a manually curated database (HPRD) that covers protein-protein interaction.

Protein-Protein Interaction Corpus

The dataset used here is the SPIES corpus for protein-protein interaction [22] which contains 963 sentences and 1436 interactions. Some examples of interactions from the dataset include (each interaction has tab-delimited four parts: the type of interaction, the first participant, the second participant and the sentence itself):

- *interaction Skp1 Fwd2 A coimmunoprecipitation assay has revealed the in vivo interaction between Skp1 and Fwd2 through the F-box domain.*
- *interact Apg3p Agp5p A cross-linking experiment revealed that Apg3p interacts with the endogenous Apg12p/Apg5p conjugate.*
- *interact Apg3p Apg12p A cross-linking experiment revealed that Apg3p interacts with the endogenous Apg12p/Apg5p conjugate.*

Since PolySearch queries are "Given X Find Associated Y", the first participant is used as the value for X. Additionally, in order to develop and test PolySearch's rule based pattern recognition system independently of the order of the participants, we expanded the corpus such that for each given pair of interactions and sentences, each participant can be used as the given X. For example:

- *interaction Skp1 Fwd2 A coimmunoprecipitation assay has revealed the in vivo interaction between Skp1 and Fwd2 through the F-box domain.*

becomes:

- *interaction Skp1 Fwd2 A coimmunoprecipitation assay has revealed the in vivo interaction between Skp1 and Fwd2 through the F-box domain.*
- *interaction Fwd2 Skp1 A coimmunoprecipitation assay has revealed the in vivo interaction between Skp1 and Fwd2 through the F-box domain.*

With the corpus expanded, the first 200 sentences of the corpus were used as a training set and the rest of the sentences were used as the test set. For association words, we first manually compiled a list of most likely protein-protein interaction words and then assembled a complete list using sentence analysis from unrelated protein-protein interaction texts. One method for the sentence analysis and extraction of association words was described in the previous chapter. The default association word list for all possible PolySearch queries can be found at http://wishart.biology.ualberta.ca/polysearch/help/association_word_list.htm. When given a protein X and a sentence, PolySeach attempts to find the protein-protein interaction(s) described in the sentence for the given protein X. Table 5.3 shows the performances of PolySearch's rule based pattern recognition system on the expanded corpus (i.e. the cut-off score is R1 >= 1).

	PolySearch R1 >= 1 (training set)	**PolySearch R1 >= 1 (test set)**	**PolySearch R2, R3 >= 1 Baseline (test set)**	**Naïve Bayes (test set)**
Precision (%)	70.0	71.1	49.1	49.4
Recall (%)	75.7	71.8	84.4	84.4
F-measure (%)	72.7 (±5.2)	71.5 (±6.5)	62.1 (±4.8)	62.3 (±4.6)

Table 5.3: Precision, recall and f-measure on a corpus of protein-protein interaction using PolySearch's rule based pattern recognition system (when X is given) and using a Naïve Bayes classifier available from Weka.

This evaluation shows that the R1 sentence and the pattern recognition approach has better performance than the baseline co-occurrence approach (i.e. a sentence that mentions two genes without an association word and without a pattern, which is what we call an R3 sentence). In this case, since each sentence in the corpus has at least one positive protein-protein interaction, each sentence also has an association word; therefore, a co-occurrence sentence with a association word (or an R2 sentence) has the same performance. In practice, not all co-occurrence sentences are R2 sentences. Therefore, the performance of a simple word co-occurrence approach could be worse with actual abstracts. Nonetheless, the word co-occurrence approach is part of most text mining systems. As shown in Table 5.3, the co-occurrence approach is a low precision approach and because of that, a co-occurrence sentence is generally less relevant. In comparison, PolySearch's combined approach has higher precision and sufficiently high recall such that the f-measure value for PolySearch is better than a simple word co-occurrence approach. The numbers in Table 5.3 mean that the sentences that PolySearch extracts will generally be more relevant while maintaining sufficient coverage.

The motivation behinds PolySearch's pattern recognition system can be seen in the following example. The sentence "Apg12p is then transferred to Apg10p, an E2-like enzyme, and conjugated with Apg5p, whereas Apg8p is transferred to Apg3p, another E2-like enzyme, followed by conjugation with phosphatidylethanolamine" has five genes mentioned: Apg3p, Apg5p, Apg8p, Apg10p, and Apg12p. Using the word co-occurrence approach, 10 possible protein-protein interactions would be predicted, when in fact the sentence only provides evidence for 3 protein-protein interactions: Apg3p-Apg8p, Apg5p-Apg12p, and Apg10p-Apg12p. A good rule to prevent overzealous predictions by the co-occurrence approach is to break the sentence into two subclauses at the word "whereas". This would eliminate six of the possible incorrect predictions resulting in a significant increase in precision. In this case, the recall would not be affected. However, not all cases are this simple and not all rules work for every single case especially when trying to further improve the precision by predicting two out of the three possible protein-protein interactions in "Apg12p is then transferred to Apg10p, an E2-like enzyme, and conjugated with Apg5p."

As Table 5.3 shows, PolySearch's R1 approach improves precision but diminishes recall in comparison to the co-occurrence approach. Nonetheless, it is also worth mentioning that a simple co-occurrence sentence (or an R3 sentence) is still scored by PolySearch and if there are no R1 and R2 sentences then PolySearch will still use R3 sentences as the key sentences to ensure proper coverage. Table 5.3 also shows the performance of an approach that used a Naïve Bayes classifier available from Weka. Weka is a data mining application written in Java [37]. The features used to build the classifier were the distances between the positions of the two genes and the association word. As seen from the numbers in this table, PolySearch's heuristic R1 association word pattern recognition system still performed better than this particular machine learning method on this data. ANOVA test showed that the performance gain from PolySearch's R1 system compared to the co-occurrence baseline approach or the Naïve Bayes approach is statistically significant.

Protein-Protein Interaction Tools Comparison

For this particular assessment, we compared PolySearch against EBIMed, IHOP, and HPRD. The goal of this assessment was to see how well each of the text mining systems could assist in the manual extraction of relevant protein-protein interaction information. This is essentially an attempt to evaluate how each of the text mining systems carries out its key/relevance sentences extraction and how the key/relevance sentences are presented to their users. The other goal of this assessment was to see how each of the text mining systems performed compared to a manually curated database. There are some advantages and disadvantages a text mining system has over a manually curated database and these will be discussed later on. For this assessment, we randomly chose five proteins (Table 5.4) and used them as input to different text mining systems to extract protein-protein interaction information.

SwissProt ID	Gene Symbol	Gene Name
Q9Y6F9	WNT6	wingless-type MMTV integration site family, member 6
Q9BZ72	PITPNM2	phosphatidylinositol transfer protein, membrane-associated 2
O60282	KIF5C	kinesin family member 5C
O60749	SNX2	sorting nexin 2
O75618	DEDD	death effector domain containing protein

Table 5.4: The five SwissProt IDs, gene symbols, and gene names randomly chosen as input to evaluate how different tools perform for protein-protein interaction.

To construct a gold standard set of protein-protein interactions for the five proteins, we looked at all the results that each of EBIMed, IHOP, and PolySearch R1 pattern recognition system returned (for EBIMed and IHOP, the species category was set to Homo sapiens) and used the key sentences that each of them extracted to determine whether the results returned by the text mining systems were false positives or true positives. If the key sentences that a text mining system returned did not satisfactorily convince us that such a protein-protein interaction existed, then it was considered a false positive. After this process, the true associations found using EBIMed, IHOP, and PolySearch were manually merged together and then combined with the list of protein-protein interactions found in HPRD for the five proteins. This "gold standard" set of protein-protein interactions was then used to evaluate the different tools in terms of precision, recall, and f-measure as shown in Table 5.5.

	HPRD	EBIMed	IHOP	PolySearch R1 >=1	PolySearch + HPRD R1 >=1
Precision	31/31 = 100%	23/83 = 27.7%	39/80 = 48.8%	64/86 = 74.4%	82/104 = 78.8%
Recall	31/99 = 31.3%	23/99 = 23.2%	39/99 = 39.4%	64/99 = 69.2	82/99 = 82.8%
F-measure	47.7 (±25.7)%	25.3 (±22.4)%	43.6 (±11.6)%	69.2 (±10.0)%	80.8 (±6.8)%

Table 5.5: Precision, recall, and f-measure for protein-protein interaction evaluation among the different tools.

For this assessment, the average number of abstracts extracted for each of the five proteins was around 90. As seen in Table 5.5, PolySearch appears to be the best tool for the extraction of protein-protein interaction information as it achieves the highest f-measure among the four different tools. PolySearch also attains the highest precision and recall among the three text mining tools with only HPRD's precision being higher than PolySearch. The advantage of using a manually curated database over a text mining tool is that the data in the manually curated database is usually of very high quality and therefore its precision is high. However, manual curation takes time and this time factor tends to limit the coverage (or recall) for many databases. In contrast to a high precision, low recall system like a manually curated database, text mining systems tend to have lower precision but higher recall. Therefore, the best approach to find the most information from a text mining tool is to allow human experts to read through the results and to extract the relevant information themselves. Indeed, one of the main goals for text mining systems such as EBIMed, IHOP, and PolySearch is to support human experts in extracting relevant information. Based on the data presented here PolySearch appears to perform the best with regard to this goal. The higher precision and recall for PolySearch shown in Table 5.5 means that users of PolySearch can find relevant information faster as the key sentences that PolySearch extracted are generally more accurate. Users can also find more information in less time because PolySearch's results (as seen in this example) covered a greater number of protein-protein interactions.

Evidently, one of the reasons why EBIMed performed poorly was that EBIMed lacks a feature for query synonym expansion to search for genes. For example, the gene "KIF5C" is also known as "KINN", "NKHC", and "NKHC2" (among other names). Both PolySearch and IHOP can carry out an automatic query synonym expansion search (i.e. querying "KIF5C", "KINN", "NKHC", "NKHC2" and other synonyms of "KIF5C" simultaneously). EBIMed cannot do this, i.e. querying "KIF5C" without its synonyms. During our evaluations, we found that EBIMed generally returned the fewest results with its lack of query synonym expansion search being one of the main reasons. In addition, one of the reasons for EBIMed's lower precision was that many of the key sentences that EBIMed returned were essentially irrelevant and did not satisfactorily convey any evidence for protein-protein interactions. For IHOP, one of the reasons contributing to its limited performance arose from the fact that IHOP's display lacks a clear organization. Furthermore, IHOP limits the number of sentences that it displays. For instance, IHOP can report that one gene co-occurs with the query gene in twelve sentences and yet only one sentence is displayed to the user. It seems that IHOP only displays sentences that meet a certain criteria. While this may be a way to limit the number of sentences and to reduce the amount of information that a user has to read, if the sentences that IHOP chooses not to display offer better evidence of an association, then this likely limits IHOP's effectiveness.

Table 5.5 also shows the performance of PolySearch with both PubMed and HPRD access turned on. One of our hypotheses is that by integrating high quality, curated biomedical databases, we can improve the coverage and precision of data retrieval. Indeed as Table 5.5 shows, both precision and recall of PolySearch improves with the inclusion of the HPRD. Mining the data in high quality, manually curated databases such as HPRD does not suffer from the difficulty of parsing free form texts and the complexity of interpreting biomedical language that makes text mining abstracts such a difficult task. As a result, mining high quality, manually curated databases is easier and more effective in retrieving correct or relevant associations.

Overall, PolySearch returns far more sentences than either IHOP or EBIMed. Furthermore, the sentences that PolySearch returns are more relevant and are ranked and displayed using the PolySearch Relevancy Index (PRI) system. PolySearch's R1 and R2 sentences generally provide more relevant information and make manual validation much easier. In summary, the use of an extensive thesaurus, in combination with association words and a sentence pattern recognition system allowed PolySearch to significantly outperform both EBIMed and IHOP in this assessment. With the integration of HPRD data into

PolySearch's text mining system, the performance advantage was even greater. Using ANOVA test, we found that the difference in f-measure between PolySearch + HPRD R1 and the other tools (IHOP, EBIMed, and HPRD) is statistically significant.

5.1.4 Evaluation of Drug/Gene Associations

For the fourth assessment, we evaluated "Given Drug Find Associated Gene" queries. The intent of this query is to find all genes that are affected or acted on by a drug. This assessment differs from the last assessment in that the queries for this assessment returned an average of 1,400 abstracts per query instead of 90 abstracts per query in the last assessment. For a text mining system to be useful, it must perform well with large amounts of data such as that used in this assessment. For this assessment, we compared PolySearch's results to EBIMed, LitMiner, and a manually curated database on drug - protein interactions, called DrugBank. DrugBank is one of the largest and most comprehensive drug and drug target databases available [6]. In particular, it contains extensive information about drug and gene/protein associations (i.e. drug metabolizing enzymes and drug targets). For this assessment, the following ten drugs were randomly chosen from DrugBank for analysis and the results are shown in Table 5.7:

DrugBank ID	Common Name
APRD00028	Tramadol
APRD00108	Pefloxacin
APRD00128	Tizanidine
APRD00136	Quinidine
APRD00294	Bumetanide
APRD00319	Fenfluramine
APRD00454	Cisapride
APRD00600	Famciclovir
APRD00706	Nizatidine
APRD00761	Dicumarol

Table 5.6: The DrugBank IDs and common names for the ten drugs randomly chosen from DrugBank for evaluating "Given Drug Find Associated Gene" queries.

	DrugBank	LitMiner	EBIMed	PolySearch R1 >= 1	PolySearch + DrugBank R1 >= 1
Precision	19/19 = 100%	24/41 = 58.5%	118/186 = 63.4%	220/358 = 61.5%	223/363 = 61.4%
Recall	19/227 = 7.9%	24/227 = 10.6%	118/227 = 52.0%	220/227 = 96.9%	223/227 = 98.2%
F-measure	15.4 (±13.7)%	17.9 (±11.8)%	57.1 (±17.7)%	75.2 (±9.5)%	75.6 (±9.2)%

Table 5.7: "Given Drug Find Associated Gene": comparing DrugBank, LitMiner, EBIMed, PolySearch with PubMed, and PolySearch with PubMed + DrugBank.

To assess PolySearch's performance, PolySearch's "Given Drug Find Associated Gene" query was run for each of the ten drugs using their common names as well as their synonyms (which were automatically generated by PolySearch). We used PolySearch in two modes. In one mode, the search was limited to PubMed abstracts only and in the second mode, we turned on PolySearch's access to DrugBank to see if this would help to improve performance. The default PolySearch settings were used in this assessment. The association word list used in this assessment contains a list consisting of most likely protein interaction words (similar to the list described in the third assessment). We only looked at the results that PolySearch's R1 system returned and then tried to map the results that EBIMed and LitMiner returned to the results that PolySearch's R1 system returned or the results derived from DrugBank. Also, based on previous observations, we chose to ignore gene names that were three letters or less. All the PolySearch extracted drug-gene associations that satisfied the previously mentioned criteria were manually verified by reading the abstracts and checking appropriate databases. All manually verified drug-gene associations including the pre-existing drug-gene associations in DrugBank were combined to derive a list of gold standard drug-gene associations. This list was used to tabulate the performance measures seen in Table 5.7.

As Table 5.7 shows, PolySearch was able to identify significantly more drug-gene interactions (and potential gene targets) than what is provided by DrugBank. On average, PolySearch found 20.9 new drug-gene associations for each of the ten drugs. The reason for this discrepancy lies in the fact that the drug targets in DrugBank are typically primary drug targets, meaning they are responsible for the therapeutic effects of many drugs, while PolySearch identified secondary drug targets in addition to primary drug targets. These secondary drug targets, which could be responsible of the side effects of drugs, can be just as important as the primary drug targets. Inclusion of these secondary drug targets into DrugBank would likely improve the coverage and utility of this database.

Table 5.7 also shows that PolySearch outperformed both EBIMed and LitMiner in this task. ANOVA test showed that the difference between PolySearch and the other tools is statistically significant. It is worth mentioning that LitMiner precomputes its results and even though this has the advantage of providing the results almost instantaneously, the precomputed results contain a shorter list of genes. Furthermore, precomputed results for some of the drugs were not available. As a result, LitMiner's performance suffered. However, even if we only looked at the set of drugs for which LitMiner's precomputed results were available, LitMiner would still perform the worst. For PolySearch with access to DrugBank turned on, the performance improved slightly over PolySearch with PubMed only and R1 >= 1. In this case, we were able to extract most of the drug-gene associations with PubMed alone. As a result, the recall improved only slightly with the DrugBank integration turned on. Nonetheless, it was still an improvement and the results from the DrugBank integration reaffirmed the results found from PubMed. Looking at the recall score of PolySearch + DrugBank R1 >= 1, we missed four drug-gene associations. This was due to the fact that those drug-gene associations were derived from DrugBank drug targets where the targets were bacterial and viral proteins.

While PolySearch did find more drug-gene interactions, higher precision is still desirable. Below we took a closer look at the precision scores for different R1 cut-off values and the precision scores for all R1 sentences found in the ten queries.

	R1 >= 1	R1 >= 2	R1 >= 3	All R1 sentences
Precision	220/358 = 61.4%	120/164 = 73.2%	90/105 = 85.7%	1148/1283 = 89.5%

Table 5.8: Precision for drug-gene associations of the ten "Given Drug Find Associated Gene" queries at different R1 scores and precision for all R1 sentences of the ten queries.

Table 5.8 shows that with an R1 score >= 1, 61.4% of the extracted associations are accurate, with an R1 score >= 2, 73.2% of the extracted associations are accurate, and with an R1 score >= 3, 85.7% of the extracted associations are accurate. It is interesting to see that R1 >= 3 achieved high precision while still finding new drug-gene associations (~ 7 new associations per drug). It seems that an R1 >= 3 can be used as a reasonable cut-off score for automatic information extraction as it achieves high precision while maintaining good coverage. With larger amounts of data, there will inevitably be more noise and therefore, the false positive rate will tend to increase. Each R1 sentence usually represents strong biomedical evidence of an association and the values of R1 >= 2 or R1 >= 3 essentially mean that repeated evidence of an association has been found, thus ensuring higher precision. While recall suffers with R1 >= 2 or R3 >= 3, the use of higher R1 cut-off scores provides a means to ensure high precision while maintaining good coverage. This kind of performance is often difficult to achieve with text mining systems employing a statistical scoring scheme. This is because a strong statistical score may or may not correlate with biomedical significance. We believe the flexibility and ease of use of the PolySearch scoring system is one of its stronger and more unique features.

In total, there were 1283 R1 sentences found in the ten "Given Drug Find Associated Gene" queries. Assuming that all R1 sentences for true associations are relevant, then the R1 sentences achieved a precision of 89.5% as 1148 of the 1283 R1 sentences were judged to be relevant. To compare this to a baseline consider this: if a user had to read all 9816 sentences in the 954 abstracts that mention the 1148 relevant R1 sentences, this would equate to a baseline precision of 11.7%. Furthermore, assuming it takes 30 seconds for a skilled individual to process an abstract this would translate to 8 hours (954 * 30 seconds) of continuous reading. The time taken by PolySearch for not only identifying the 1283 R1 sentences in the 954 abstracts but also searching through the 14,000 total extracted abstracts was about 20 minutes. This shows that using PolySearch is significantly faster than using PubMed and manually reading abstracts. A similar type of analysis, with different queries, was done by the authors of EBIMed when they evaluated their system [13]. These authors reported that EBIMed achieved 39% precision in extracting relevant sentences compared to a baseline precision of 13%. Though not a direct comparison, PolySearch still appears to be a better system in extracting relevant sentences (89.5% precision versus 39% precision).

Overall with this assessment we demonstrated that PolySearch can serve as an automatic information extraction system for large amounts of biomedical data. We also demonstrated that PolySearch outperformed other text mining tools and that it can extract more data than is contained in a manually curated database. The precision that PolySearch achieved is significantly better than using PubMed and manually reading abstracts. By varying the R1 cut-off scores, PolySearch can be an information extraction system with high recall and moderate precision (R1 >= 1) or a system with high precision and moderate recall (R1 >= 3) that is still capable of finding new associations not found in a high quality manually curated database.

5.1.5 Evaluation of Metabolite/Gene Associations

In the fifth assessment, we evaluated "Given Metabolite Find Associated Gene" queries. With this assessment, we investigated how PolySearch performs with another type of automatic information extraction task. For this assessment, we compared PolySearch, EBIMed, and LitMiner to the Human Metabolome Database (HMDB). The HMDB is a database containing detailed chemical, biological and clinical information about small molecule metabolites found in the human body [7]. The HMDB also contains metabolic enzyme data for each of the metabolites. These metabolite/metabolic enzyme associations are the ones that we are interested in and what we wish to compare to PolySearch's results for its "Given Metabolite Find Associated Gene" queries. To make this assessment, the following ten metabolites were randomly chosen from the HMDB:

HMDB ID	Common Name
HMDB00210	Pantothenic acid
HMDB00721	Glycyl-L-proline
HMDB00835	N-Acetyl-a-D-galactosamine
HMDB01059	1D-Myo-inositol 1,3,4,5-tetrakisphosphate
HMDB01175	Malonyl-CoA
HMDB01381	Prostaglandin H2
HMDB01413	Citicoline
HMDB01489	Ribose 1-phosphate
HMDB01550	S-Formylglutathione
HMDB02037	12-Hydroxyeicosatetraenoic acid

Table 5.9: The HMDB IDs and common names for the ten metabolites randomly chosen from HMDB for evaluating Given Metabolite Find Associated Gene queries.

PolySearch's "Given Metabolite Find Associated Gene" query was run for each of the ten metabolites using their common names as well as their synonyms. This time PubMed, OMIM, and HMDB were used as data sources for the search. Other search settings were the same as mentioned in the Drug-Gene assessment. The average number of abstracts identified per query was 880. A list of gold standard metabolite-gene associations for these ten metabolites was compiled by manually verifying PolySearch metabolite-gene associations for R1 >= 1 and by manually compiling data from the HMDB. Table 5.10 shows a comparison between HMDB, EBIMed, LitMiner, PolySearch with PubMed alone, PolySearch with PubMed and OMIM, and PolySearch with PubMed, OMIM and HMDB.

	HMDB	LitMiner	EBIMed	PolySearch R1 >= 1	PolySearch + OMIM R1 >= 1	PolySearch + OMIM + HMDB R1 >=1
Precision	26/26 = 100%	5/5 = 100%	83/111 = 74.8%	166/263 = 63.1%	170/267 = 63.7%	183/284 = 64.4%
Recall	26/183 = 14.2%	5/183 = 2.7%	83/183 = 45.4%	166/183 = 90.7%	170/183 = 92.9%	183/183 = 100%
F-measure	24.9 (±17.6)%	5.3 (±9.3)%	56.5 (±28.4)%	74.4 (±8.7)%	75.6 (±8.1)%	78.4 (±7.6)%

Table 5.10: "Given Metabolite Find Associated Gene": precision, recall and f-measure for HMDB, LitMiner, EBIMed, PolySearch with PubMed, PolySearch with PubMed + OMIM, and PolySearch with PubMed + OMIM + HMDB.

Overall, PolySearch appears to be just as effective in automatic information extraction for metabolite-gene associations as it is for drug-gene associations. On average, PolySearch found 15.7 new metabolite-gene associations for each of the ten metabolites. Table 5.10 also shows that the performance of PolySearch using PubMed + OMIM + HMDB is the best. This assessment demonstrates again that mining of high quality, manually curated databases can help improve both the sensitivity and specificity of biomedical information extraction. ANOVA test showed that PolySearch is statistically significantly better than the other tools.

Next, we took a closer look at the precision scores for different R1 score cut-offs and the precision scores for all R1 sentences to see if high precision with moderate recall could be achieved.

	R1 >= 1	R1 >= 2	R1 >= 3	All R1 sentences
Precision	166/263 = 63.1%	85/105 = 81.0%	54/60 = 90.0%	828/926 = 89.4%

Table 5.11: Precision for metabolite-gene associations of the ten "Given Metabolite Find Associated Gene" queries at different R1 scores and precision for all R1 sentences of the ten queries.

An R1 cut-off score of >= 3 achieved a precision of 90% while still finding new metabolite-gene associations (~ 4 new associations per metabolite). This suggests that PolySearch can be tuned to achieve high precision and moderate recall which means that it could be used for automated text mining. As noted on Table 5.11, PolySearch achieved 89.4% precision using R1 sentences. Compared to the PubMed baseline, to manually retrieve the 828 relevant R1 sentences from the 740 abstracts (7287 sentences) equates to an 11.4% precision. Again, if we assume that it takes an individual ~30 seconds to process an abstract, this manual search would take around 6 hours. The time taken by PolySearch to search through the 8800 total extracted abstracts and identify the 926 R1 sentences in the 740 abstracts was about 15 minutes. Clearly PolySearch is significantly faster than manually reading PubMed abstracts alone.

5.1.6 Evaluation of Disease/Gene Associations

For the sixth assessment, we evaluated "Given Disease Find Associated Gene" queries. This is a relatively common query supported by a number of text mining tools. For this assessment, we compared PolySearch, EBIMed, and LitMiner to the Genetic Association Database (GAD). GAD is a manually curated archive of human genetic association studies of complex diseases and disorders [27] and it is also one of the databases that has been integrated into PolySearch. The ten diseases shown in Table 5.12 were chosen as queries since they represent a mixture of monogenetic diseases and complex genetic disorders.

Disease Name
Alkaptonuria
Cylindromatosis
Gilbert syndrome
McLeod syndrome
Motor neuron disease
Omphalocele
Onchocerciasis
Orofacial cleft
Synpolydactyly
Vitelliform macular dystrophy

Table 5.12: The disease names for the ten diseases randomly chosen for evaluating Given Disease Find Associated Gene queries

For this assessment, we carried out a complete evaluation using various PolySearch options. We looked at a PolySearch cut-off score of R2 >= 1 and a cut-off score of R1 >= 1 to see if PolySearch's R1 pattern recognition system could improve performance. We also looked at the results with access to the OMIM and GAD database turned on, as both of these databases contain a good deal of disease-gene information. The average number of abstracts available per disease query was 733. The list of gold standard disease-gene associations was compiled using the disease-gene associations listed in GAD as well as manually verified true disease-gene associations of R1 >= 1 or R2 >= 1. Table 5.13 shows the performance of the different text mining tools.

	GAD	LitMiner	EBIMed	PolySearch R2 >= 1	PolySearch R1 >= 1	PolySearch + OMIM R1 >= 1	PolySearch + OMIM + GAD R1 >= 1
Precision	21/21 = 100%	4/5 = 80%	102/177 = 57.8%	119/251 = 47.4%	93/133 = 69.9%	101/143 = 70.6%	113/156 = 72.4%
Recall	21/132 = 15.9%	4/132 = 3.0%	102/132 = 77.3%	119/132 = 90.2%	93/132 = 70.4%	101/132 = 76.5%	113/132 = 85.6%
F-measure	27.5 (±23.0)%	5.8 (±13.5)%	66.0 (±10.3)%	62.1 (±16.6)%	70.2 (±17.5)%	73.5 (±9.3)%	78.5 (±10.3)%

Table 5.13: "Given Disease Find Associated Gene": precision, recall and f-measure for GAD, LitMiner, EBIMed, PolySearch R2 >= 1, PolySearch R1 >= 1, PolySearch with PubMed + OMIM, and PolySearch with PubMed + OMIM + GAD.

Generally speaking, PolySearch alone performed the best among the tools in this assessment while PolySearch + OMIM + GAD achieved the best f-measure value. PolySearch R1 >= 1 performed better than PolySearch R2 >= 1 and we also observed a steady increase in performance with PolySearch OMIM integration turned on and PolySearch OMIM + GAD integrations turned on. Taken together, these findings confirm our hypotheses that PolySearch's R1 pattern recognition schema and PolySearch's integration of manually curated databases into its search protocol can improve a text mining system's ability to extract associations. This assessment also showed that PolySearch outperformed both LitMiner and EBIMed in extracting disease-gene associations. ANOVA showed that both PolySearch and EBIMed are statistically significantly better than LitMiner and GAD; however, there is no statistically significant difference between PolySearch and EBIMed. In this assessment, PolySearch found an average of 11.1 new disease-gene associations (not listed in GAD) for each of the ten diseases.

	R1 >= 1	R1 >= 2	R1 >= 3	All R1 sentences
Precision	93/133 = 69.9%	39/45 = 86.7%	27/28 = 96.4%	394/430 = 91.6%

Table 5.14: Precision for disease-gene associations of the ten "Given Disease Find Associated Gene" queries at different R1 scores and precision for all R1 sentences of the ten queries

As shown in Table 5.14, an R1 cut-off score >= 1 achieved a moderate precision of around 70%. An R1 cut-off score >= 2 allowed PolySearch to achieve a precision of 86.7% while still finding approximately 2 new disease-gene associations per disease. With regard to all R1 sentences, 394 out of 430 R1 sentences were deemed relevant, resulting in a 91.6% precision for R1 sentences. Compared to the PubMed baseline, manually retrieving the 394 R1 sentences out from the 352 abstracts (3163 sentences) equates to a 12.5% precision. Again, if we assume that it takes an individual ~30 seconds to process an abstract, this manual search would take around 3 hours. The time taken by PolySearch to search through the 7300 total extracted abstracts and identify the 430 R1 sentences in the 352 abstracts was about 10 minutes. Once again, this demonstrates how PolySearch provides a significant time savings compared to using PubMed and manually reading abstracts alone.

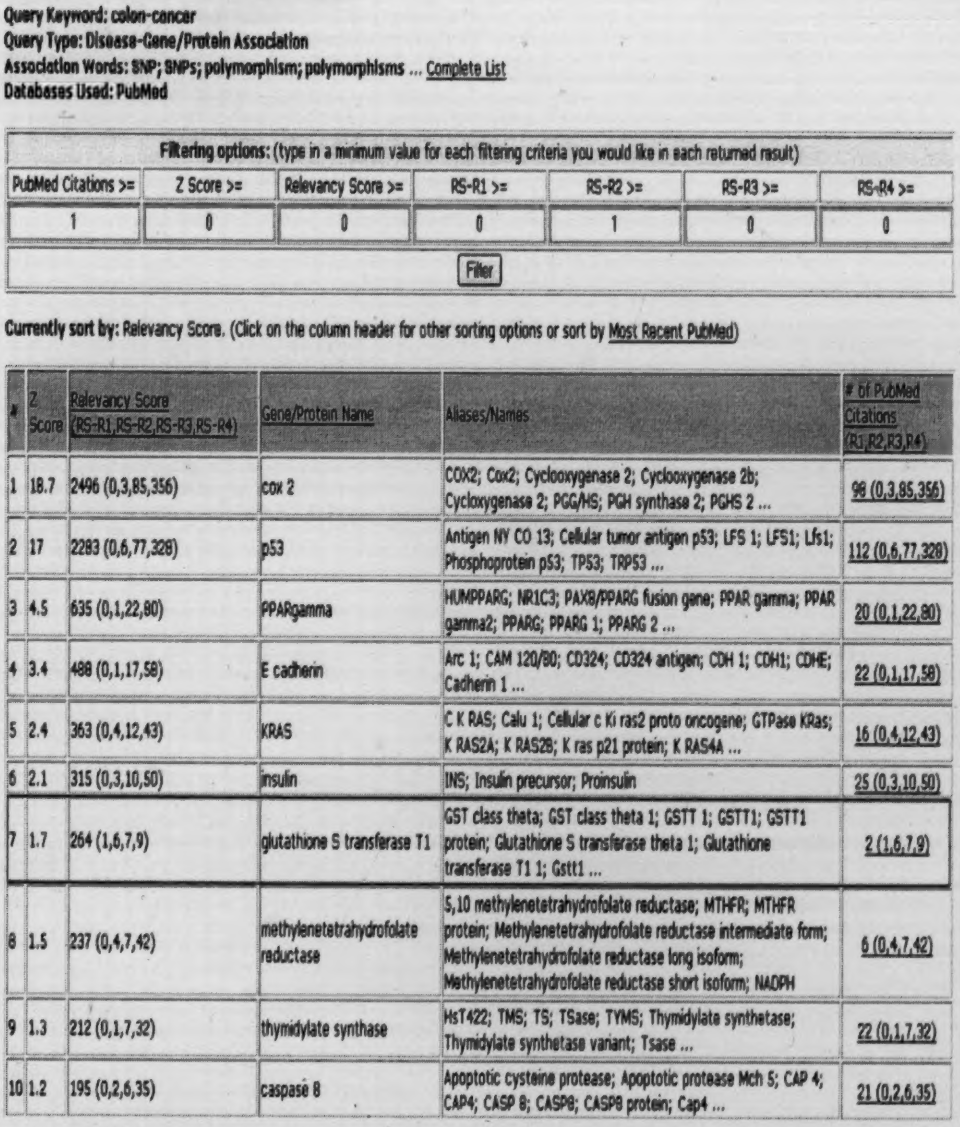
5.1.7 Manual versus Automated

For the seventh assessment, PolySearch was compared with the speed/coverage performance of a researcher tasked with finding all metabolites known to be in human cerebrospinal fluid and obtaining their concentration values. The individual (a senior undergraduate, now in medical school) was given 6 weeks and several primary references (PubMed, the Merck Manual [38], Google Scholar [39], Wikipedia [40]) to assist with his search. The student was encouraged to access and read through abstracts, complete journal articles and clinical chemistry reference books to obtain the necessary data. Additionally the student's progress (i.e. metabolite list) was tracked on a weekly basis and continuous suggestions were provided to improve his search methodology. After the student had completed his 6 week search project, the number of metabolite concentration values identified by the student using manual methods was just 47 (in 42 days). PolySearch then was run using the "Given Text Word Find Metabolites" query, with the text words being "CSF" and "cerebrospinal fluid" and the resulting list was given to the student to help his search. At the end of the student's search project (roughly 9 weeks later), a total of 308 concentration values were reported by the student with 70% of these being obtained through PolySearch. The student indicated that he considered PolySearch to be far a better search tool than his manual approach and that PolySearch had helped him tremendously. While it may be argued that such an assessment lacks the scientific rigour found in our other evaluations, we believe these results are perhaps the most realistic in terms of demonstrating the potential time-savings and the breadth of coverage that are possible with a robust text-mining system.

5.1.8 Final Example

For our last example, we will illustrate how all of PolySearch's functionalities can be used to go from a given disease to identify specific SNPs associated with the disease gene to the design of specific PCR primers.

SNPs or single nucleotide polymorphisms account for about 90% of all human genetic variation. They occur every 100 to 300 bases along the 3-billion-base human genome. It is generally believed that a better understanding of SNPs could lead to better approaches to treating or diagnosing diseases [41]. To use PolySearch to study SNPs, one can just use "SNP, SNPs, Polymorphism, Polymorphisms" as the association words. Figure 5.1 shows a "Disease-to-Gene" query where the query is "colon cancer", the association words are "SNP", "SNPs", "Polymorphism" and "Polymorphisms" and the minimum R2 filter is set to 1 leaving only the results that are highly relevant to the association words associated with SNPs.



Query Keyword: colon-cancer
Query Type: Disease-Gene/Protein Association
Association Words: SNP; SNPs; polymorphism; polymorphisms ... Complete List
Databases Used: PubMed

Filtering options: (type in a minimum value for each filtering criteria you would like in each returned result)

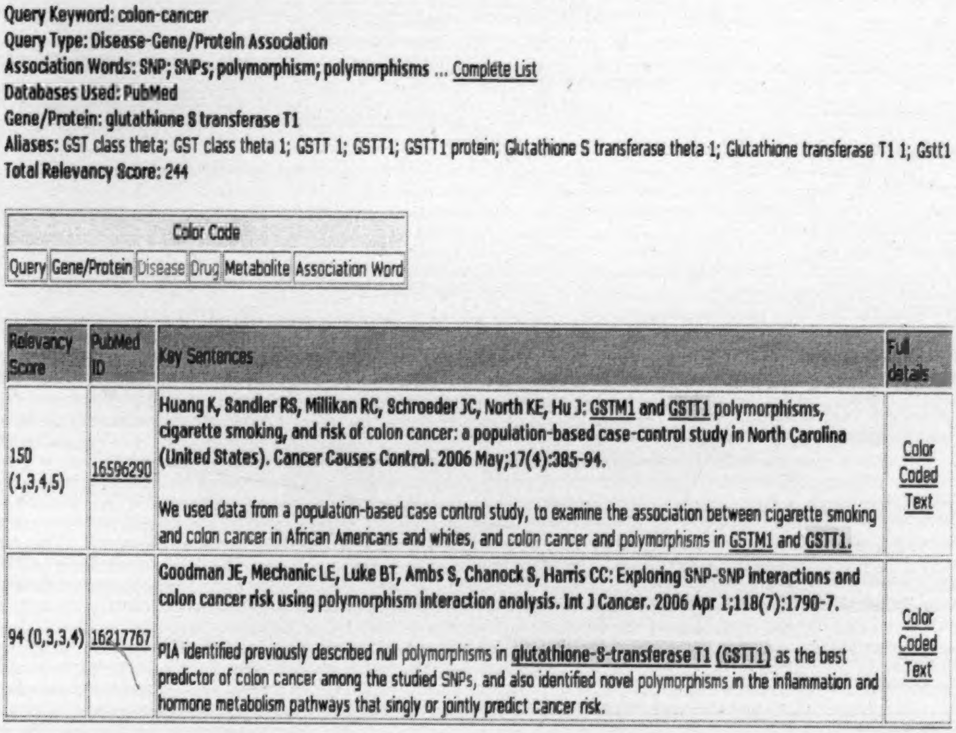
PubMed Citations >=	Z Score >=	Relevancy Score >=	RS-R1 >=	RS-R2 >=	RS-R3 >=	RS-R4 >=
1	0	0	0	1	0	0

Currently sort by: Relevancy Score. (Click on the column header for other sorting options or sort by Most Recent PubMed)

#	Z Score	Relevancy Score (RS-R1,RS-R2,RS-R3,RS-R4)	Gene/Protein Name	Aliases/Names	# of PubMed Citations (R1,R2,R3,R4)
1	18.7	2496 (0,3,85,356)	cox 2	COX2; Cox2; Cyclooxygenase 2; Cyclooxygenase 2b; Cycloxygenase 2; PGG/HS; PGH synthase 2; PGHS 2 ...	98 (0,3,85,356)
2	17	2283 (0,6,77,328)	p53	Antigen NY CO 13; Cellular tumor antigen p53; LFS 1; LFS1; Lfs1; Phosphoprotein p53; TP53; TRP53 ...	112 (0,6,77,328)
3	4.5	635 (0,1,22,80)	PPARgamma	HUMPPARG; NR1C3; PAX8/PPARG fusion gene; PPAR gamma; PPAR gamma2; PPARG; PPARG 1; PPARG 2 ...	20 (0,1,22,80)
4	3.4	488 (0,1,17,58)	E cadherin	Arc 1; CAM 120/80; CD324; CD324 antigen; CDH 1; CDH1; CDHE; Cadherin 1 ...	22 (0,1,17,58)
5	2.4	363 (0,4,12,43)	KRAS	C K RAS; Calu 1; Cellular c Ki ras2 proto oncogene; GTPase KRas; K RAS2A; K RAS2B; K ras p21 protein; K RAS4A ...	16 (0,4,12,43)
6	2.1	315 (0,3,10,50)	insulin	INS; Insulin precursor; Proinsulin	25 (0,3,10,50)
7	1.7	264 (1,6,7,9)	glutathione S transferase T1	GST class theta; GST class theta 1; GSTT 1; GSTT1; GSTT1 protein; Glutathione S transferase theta 1; Glutathione transferase T1 1; Gstt1 ...	2 (1,6,7,9)
8	1.5	237 (0,4,7,42)	methylenetetrahydrofolate reductase	5,10 methylenetetrahydrofolate reductase; MTHFR; MTHFR protein; Methylenetetrahydrofolate reductase intermediate form; Methylenetetrahydrofolate reductase long isoform; Methylenetetrahydrofolate reductase short isoform; NADPH	6 (0,4,7,42)
9	1.3	212 (0,1,7,32)	thymidylate synthase	HsT422; TMS; TS; TSase; TYMS; Thymidylate synthetase; Thymidylate synthetase variant; Tsase ...	22 (0,1,7,32)
10	1.2	195 (0,2,6,35)	caspase 8	Apoptotic cysteine protease; Apoptotic protease Mch 5; CAP 4; CAP4; CASP 8; CASP8; CASP8 protein; Cap4 ...	21 (0,2,6,35)

Figure 5.1: Results from a "Disease to Gene" query where the query is colon cancer, the filter words are SNP, SNPs, Polymorphism and Polymorphisms, the minimum R2 filter is set to 1, and the maximum number of abstracts is set to 2000.

COX-2 is a well known gene associated with colon cancer so it is not surprising to see that some SNP studies have been done in relation to colon cancer and COX-2. By briefly scanning through the results, we see that glutathione S transferase T1 (GSTT1) stands out as having only two abstracts but GSTT1 has the highest R1 and R2 scores. Looking at the key sentences for GSTT1 (Figure 5.2), it seems that the SNP/polymorphism association of colon cancer and GSTT1 is quite strong and maybe worth further investigation. By looking at the relevancy score and the PRI (without even looking at the key sentences), one can quickly gain insight on the strength of this association.



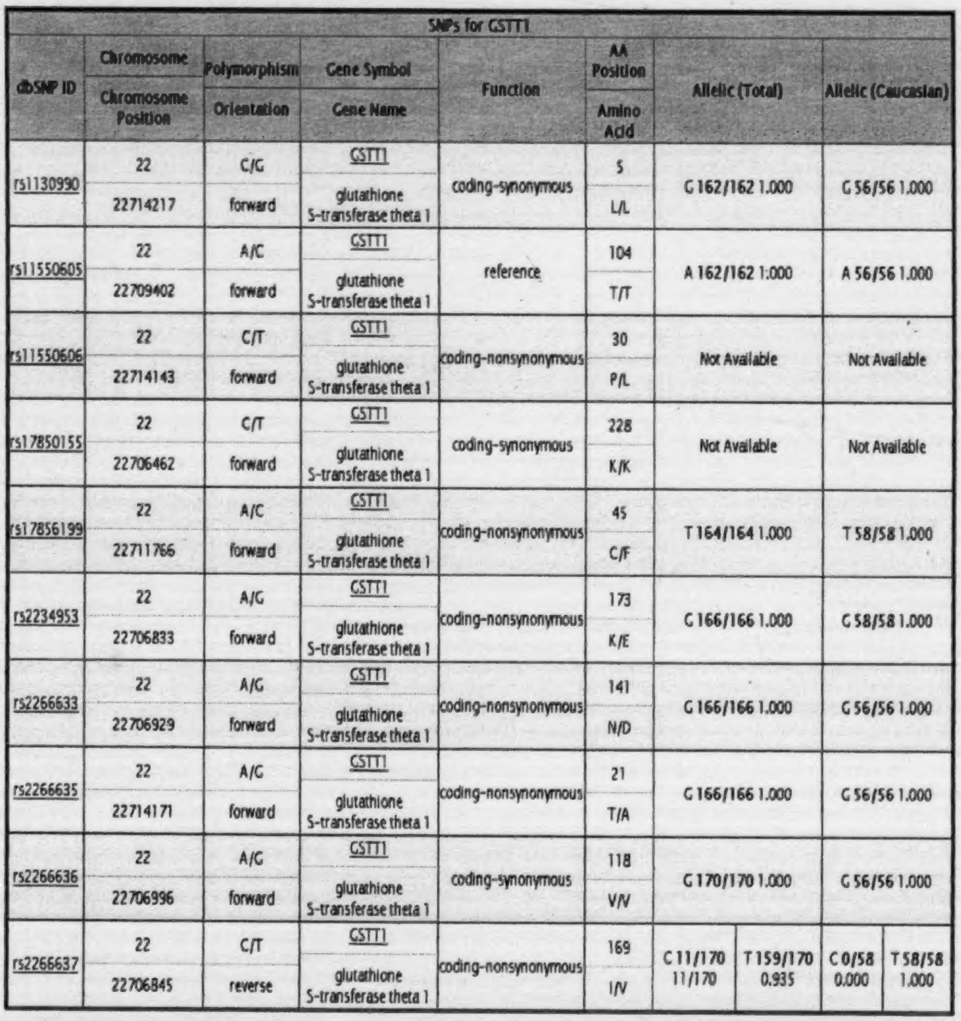
Query Keyword: colon-cancer
Query Type: Disease-Gene/Protein Association
Association Words: SNP; SNPs; polymorphism; polymorphisms ... Complete List
Databases Used: PubMed
Gene/Protein: glutathione S transferase T1
Aliases: GST class theta; GST class theta 1; GSTT 1; GSTT1; GSTT1 protein; Glutathione S transferase theta 1; Glutathione transferase T1 1; Gstt1
Total Relevancy Score: 244

Color Code					
Query	Gene/Protein	Disease	Drug	Metabolite	Association Word

Relevancy Score	PubMed ID	Key Sentences	Full details
150 (1,3,4,5)	16596290	**Huang K, Sandler RS, Millikan RC, Schroeder JC, North KE, Hu J: GSTM1 and GSTT1 polymorphisms, cigarette smoking, and risk of colon cancer: a population-based case-control study in North Carolina (United States). Cancer Causes Control. 2006 May;17(4):385-94.** We used data from a population-based case control study, to examine the association between cigarette smoking and colon cancer in African Americans and whites, and colon cancer and polymorphisms in GSTM1 and GSTT1.	Color Coded Text
94 (0,3,3,4)	16217767	**Goodman JE, Mechanic LE, Luke BT, Ambs S, Chanock S, Harris CC: Exploring SNP-SNP interactions and colon cancer risk using polymorphism interaction analysis. Int J Cancer. 2006 Apr 1;118(7):1790-7.** PIA identified previously described null polymorphisms in glutathione-S-transferase T1 (GSTT1) as the best predictor of colon cancer among the studied SNPs, and also identified novel polymorphisms in the inflammation and hormone metabolism pathways that singly or jointly predict cancer risk.	Color Coded Text

Figure 5.2: The key sentences for SNP/polymorphism association between colon cancer and GSTT1.

To further study GSTT1, one can use PolySearch's "Gene to SNP" search by entering GSTT1 as the input gene symbol. Figure 5.3 shows the PolySearch's output for "Gene-to-SNP" using GSTT1 as the input.



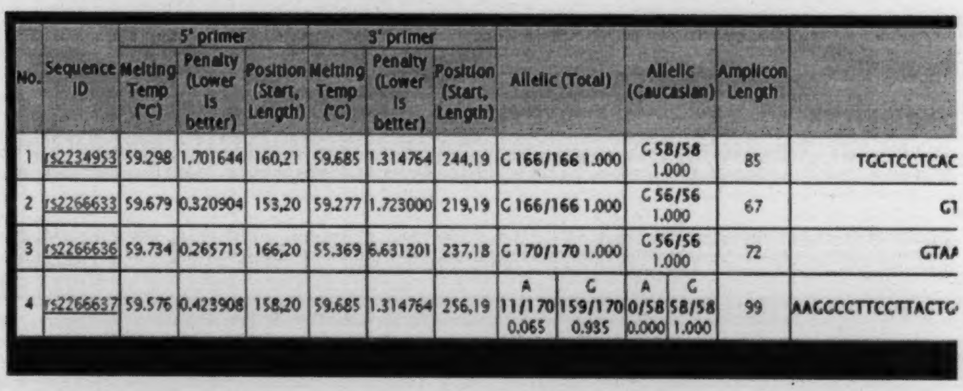
SNPs for GSTT1

dbSNP ID	Chromosome	Polymorphism	Gene Symbol	Function	AA Position	Allelic (Total)	Allelic (Caucasian)
	Chromosome Position	Orientation	Gene Name		Amino Acid		
rs1130990	22	C/G	GSTT1	coding-synonymous	5	G 162/162 1.000	G 56/56 1.000
	22714217	forward	glutathione S-transferase theta 1		L/L		
rs11550605	22	A/C	GSTT1	reference	104	A 162/162 1:000	A 56/56 1.000
	22709402	forward	glutathione S-transferase theta 1		T/T		
rs11550606	22	C/T	GSTT1	coding-nonsynonymous	30	Not Available	Not Available
	22714143	forward	glutathione S-transferase theta 1		P/L		
rs17850155	22	C/T	GSTT1	coding-synonymous	228	Not Available	Not Available
	22706462	forward	glutathione S-transferase theta 1		K/K		
rs17856199	22	A/C	GSTT1	coding-nonsynonymous	45	T 164/164 1.000	T 58/58 1.000
	22711766	forward	glutathione S-transferase theta 1		C/F		
rs2234953	22	A/G	GSTT1	coding-nonsynonymous	173	G 166/166 1.000	G 58/58 1.000
	22706833	forward	glutathione S-transferase theta 1		K/E		
rs2266633	22	A/G	GSTT1	coding-nonsynonymous	141	G 166/166 1.000	G 56/56 1.000
	22706929	forward	glutathione S-transferase theta 1		N/D		
rs2266635	22	A/G	GSTT1	coding-nonsynonymous	21	G 166/166 1.000	G 56/56 1.000
	22714171	forward	glutathione S-transferase theta 1		T/A		
rs2266636	22	A/G	GSTT1	coding-synonymous	118	G 170/170 1.000	G 56/56 1.000
	22706996	forward	glutathione S-transferase theta 1		V/V		
rs2266637	22	C/T	GSTT1	coding-nonsynonymous	169	C 11/170 11/170 T 159/170 0.935	C 0/58 0.000 T 58/58 1.000
	22706845	reverse	glutathione S-transferase theta 1		I/V		

Figure 5.3: The output for a "Gene to SNP" search using GSTT1 as input. PolySearch collects important information about SNP such as: position, type of polymorphism, gene symbol/name, function and allelic frequency.

Once the "Gene-to-SNP" query has been completed, then the "SNP to PCR Primer Design" search can be used to design PCR primers for the relevant SNPs.

Figure 5.4 shows the PolySearch's output for "SNP to PCR Primer Design" using rs2234953, rs2266633, rs2266636 and rs2266637 as input.



No.	Sequence ID	5' primer			3' primer			Allelic (Total)	Allelic (Caucasian)	Amplicon Length	
		Melting Temp (°C)	Penalty (Lower is better)	Position (Start, Length)	Melting Temp (°C)	Penalty (Lower is better)	Position (Start, Length)				
1	rs2234953	59.298	1.701644	160,21	59.685	1.314764	244,19	C 166/166 1.000	G 58/58 1.000	85	TGCTCCTCAC
2	rs2266633	59.679	0.320904	153,20	59.277	1.723000	219,19	C 166/166 1.000	G 56/56 1.000	67	GT
3	rs2266636	59.734	0.265715	166,20	55.369	6.631201	237,18	G 170/170 1.000	G 56/56 1.000	72	GTA
4	rs2266637	59.576	0.423908	158,20	59.685	1.314764	256,19	A 11/170 0.065 G 159/170 0.935	A 0/58 0.000 G 58/58 1.000	99	AAGCCCTTCCTTACTG

Figure 5.4: An illustration of the PCR Primer Design feature in PolySearch.

The "SNP to PCR Primer Design" search returns: 1) 5' primer information, 2) 3' primer information, 3) allele frequencies and 4) a colour coded amplicon sequence. The primer design results can be downloaded in tab delimited format which can be opened in MS Excel. This example illustrates some of the integrated search functionalities in PolySearch. It is just one example of many possible integrated searches that PolySearch supports.

5.2 Discussion

One of the most common queries (and among the most common experiments) in biomedical research is "Given X, find all associated Y's". Examples of such queries might be "find all polymorphisms associated with breast cancer" or "find all metabolites associated with kidney transplant rejections". With the growing volume of data from genomic, metabolomic and proteomic experiments these types of queries are becoming increasingly common. However, with most existing database search tools these queries are not so easily answered. PolySearch was specifically designed to answer these kinds of associative questions. Through its use of customized thesauruses PolySearch allows users to query eight of the most common multi-component biological concepts including: diseases, genes/proteins, SNPs, drugs, metabolites, tissues, organs, and subcellular localization. In addition, users may also provide their own thesauruses to permit any types of specialized text searches.

PolySearch brings a number of useful innovations to the area of biomedical text mining. First, it expands the breadth of query possibilities by including a much more diverse set of thesauruses or synonym dictionaries. Second, it incorporates the very useful textual data found in many other biological databases (OMIM, DrugBank, SwissProt, HMDB, HPRD, GAD) and integrates this into its text mining and text scoring processes. As the evaluations show, integrating these manually curated databases improves the performance of PolySearch's text mining system by increasing both precision and recall. Third, PolySearch employs a scoring scheme that combines both word co-occurrence and sentence patterns. This leads to a relevancy score ranking that is aided by the PolySearch Relevancy Index (PRI) to help users immediately grasp the quality of any association. The PRI (in particular R1 and R2 scores) is also used to as an innovative scoring cut-off for automatic information extraction as shown in the "Given Drug Find Associated Gene", the "Given Metabolite Find Associated Gene", and the "Given Disease Find Associated Gene" assessments (sections 5.1.4, 5.1.5 and 5.1.6). In general, PolySearch exhibits the best overall performance of the text mining tools that we assessed. Furthermore, PolySearch is significantly faster than using PubMed and manually reading abstracts. Fourth, the use of customizable association words and R1, R2 sentences provides users with useful information and direct control over how to rank or assess the text-derived associations. Instead of giving users a ranking system that is pre-defined, PolySearch gives the control back to the user, allowing a user to use any association words they choose. In addition to these innovations, PolySearch also borrows a number of excellent ideas from existing text mining systems. These include the use of colour-coded word highlighting schemes (as found in MedMiner, iHOP, ALIBABA and EBIMed), the selection and display of key sentences (as found in MedMiner, iHOP, ALIBABA and EBIMed), the extensive use of hyperlinks (as found in Entrez, iHOP and EBIMed) and the connectivity to multiple databases (as found in iHOP and Entrez).

Relative to other biomedical text mining systems PolySearch appears to be unique in its ability to link text mining with SNP analyses. With the completion of the HapMap project [42] and the growing importance of pharmacogenomics and SNP profiling [43] we believe this capacity will be particularly important to many biomedical and pharmaceutical researchers. PolySearch gathers data from a wide variety of SNP and mutation databases and integrates this data to provide critical information about the potential importance of a SNP or mutation. The mutation data from HGMD [5] links mutation information to specific diseases, providing a valuable resource for disease-mutation associations. For SNP analyses, PolySearch assembles data on a variety of SNP features including: SNP function (eg: synonymous, nonsynonymous, coding, non-coding), level of validation, allele name, allele position, allele frequencies, amino acid change, and amino acid position. Based on these features, researchers can quickly select SNPs of interest for further study. For example, a SNP that is nonsynonymous, validated, and has varying allele frequencies among different ethnic populations will generally be more useful than a SNP that is synonymous and not validated.

Because of its SNP data mining capacity PolySearch can be used purely as a SNP analysis tool. For instance, users can retrieve SNP data such as function and allele frequencies and use them to determine which SNPs need specially designed primers using PolySearch's primer design tool. Alternatively, PolySearch can be used to study disease-SNP, gene-SNP, drug-SNP and metabolite-SNP associations. For example, one can use PolySearch to construct a query such as "Given drug find gene(s)" to generate a list of genes that have possible associations to the query drug. Once a gene or set of interesting genes is identified, users may then employ the "Given gene find SNP" query to find SNPs that are associated with the gene or genes of interest. Once the SNPs are identified, users may continue to use PolySearch to help carry on with SNP analysis. The example given in section 5.1.8 demonstrates how this can be done for disease-SNP associations. Similar methods can be applied to other types of SNP analyses.

PolySearch can also be used as a tool to facilitate database annotation and database curation. Similar efforts have been undertaken with LitMiner and

WikiGene. In our case we have used PolySearch to annotate the Human Metabolite Database (HMDB) [7] and DrugBank [6]. For instance, the "Given text word find metabolites" has been used to determine the presence and concentrations of metabolites in a number of biofluids. Likewise, the "Given metabolite-Find Diseases" search has been used to help identify particular metabolites or metabolite profiles associated with a variety of diseases. The feedback from annotators has been very positive, suggesting that PolySearch is a particularly useful tool for database annotation. This feedback has also helped to improve PolySearch's user interface and to expand its capabilities.

Text mining is still an imperfect science and so the results of any PolySearch query should always be treated with some prudence. Therefore users are always advised to take some time to read and analyze any PolySearch results. Indeed PolySearch is really designed to facilitate user validation. As with any text mining tool, PolySearch's results should be considered as fragmentary pieces of evidence or potential hypotheses that require some degree of intelligent scrutiny before a solid link or a definitive association can be made. Typically the most frequently cited associations or those with the highest relevancy scores can be assumed to have solid supporting evidence. Therefore whenever a R1, R2, or R3 hit to a non-PubMed database occurs, one should consider that association to be generally well-supported. When relevancy scores are low and there are no R1, R2 or R3 hits, then the validity of the association is probably dubious. With most areas concerning textual analysis, there are no hard and fast rules; however, with PolySearch we believe that we have made some important progress in terms of automatic information extraction as shown with a number of our evaluations.

Chapter 6

6. Conclusion and Future Work

6.1 Conclusion

In this thesis, we presented a brief survey of existing biomedical text mining systems and introduced a new tool, PolySearch, which offers diverse biomedical search and text ranking possibilities. As seen from the descriptions given in Chapters 4 and 5, PolySearch is a collection of tools designed for flexible and diverse text mining applications. In particular, it supports the mining and discovery of associations between diseases, genes, drugs, metabolites and SNPs. PolySearch provides a ranking scheme to assess the strength of these associations and displays its results in a transparent manner. Users can also use PolySearch to gather SNP data concerning a variety of SNP features including: SNP function, validation, allele, allele position, amino acid, amino acid position, and allele frequencies. The functionalities in PolySearch can be used independently or in an integrated manner. PolySearch thus represents a novel attempt to deal with the fragmentary, non-contextual nature of most biomedical data.

6.2 Future Work

PolySearch is not without some limitations. As a text mining tool, PolySearch uses a relatively simple dictionary approach to identify biological or biomedical entities. This means PolySearch cannot identify novel or newly named diseases, genes, cell types, drugs or metabolites. Another limitation lies in its inability to extract context or meaning from sentences or terms. Indeed, the majority of PolySearch's errors come from incorrect term identification (the identified term has a different meaning) or incorrect word associations such as "Drug X and Protein Y were used together to improve Process Z" or "Drug X and Drug W inhibit Protein Y and Protein Z respectively" (in this case the association between X and Z and the association between W and Y would be incorrect). Methods that use artificial intelligence (AI), word context or machine learning (ML) methods could potentially improve the current term identification system [44]. Likewise, AI or ML methods could potentially decipher the context or meaning of the biological/biomedical associations they find. These technologies could be used in the future to further complement PolySearch's dictionary approach and its relatively simple sentence pattern recognition system. No doubt as the individual components in PolySearch improve, the overall effectiveness of PolySearch would become better.

Currently, PolySearch employs a manual interactive relevancy feedback mechanism. While we used this manual approach to improve the selection of association words for PolySearch's default searches (thus improving PolySearch's performance), it is unknown how well novice users can take advantage this feature for more customized searches. In addition, for the "Given Text Word Find Associated Text Word" search, it is impossible for us to provide a default association word list. In the future, we would like to incorporate an automatic relevancy feedback mechanism in PolySearch such that PolySearch can automatically provide suggestions for relevant association words.

PolySearch is also somewhat limited in its information display capabilities. Currently for each query, PolySearch returns a list of results ranked by a relevancy score. For each possible association, PolySearch displays key sentences and a hyperlinked "details" view. For queries with relatively few associations, this presentation is still quite effective. However for queries with many associations an automatically generated summary would be more useful. The next step for PolySearch is to possibly organize key sentences into different categories, or to use only short and conclusive sentences similar to Chilibot [45], or possibly to use sentence clustering method similar to CIDR [46]. These auto-generated summaries would ideally display short search synopses that would link back to PolySearch's results. The auto-generated summaries could also be used as short stubs for new entries into manually curated databases.

While PolySearch provides an impressive array of tools for SNP analyses, there are a number of new and sophisticated SNP tools available such as: flanking SNP query [47], SNP flanking sequence [47], SNP cutter [47], and Pyro primer [47] that offer very useful SNP annotation features that could complement what PolySearch generate. We hope to integrate one or more these tools into PolySearch in the future as a way to further enhance PolySearch and its data mining/analysis capabilities.

Bibliography

1. Baasiri, R.A., Glasser, S.R., Steffen, D.L. and Wheeler, D.A: **The breast cancer gene database: a collaborative information resource.** Oncogene 1999. 18:7958-7965.

2. Hersh, W.R: **Information retrieval: a health and biomedical perspective.** 2nd ed. New York: Springer; 2003.

3. Wheeler, D.L., Barrett, T., Benson, D.A., Bryant, S.H., Canese, K., Church, D.M., DiCuccio, M., Edgar, R., Federhen, S., Helmberg, W., Kenton, D.L., Khovayko, O., Lipman, D.J., Madden, T.L., Maglott, D.R., Ostell, J., Pontius, J.U., Pruitt, K.D., Schuler, G.D., Schriml, L.M., Sequeira, E., Sherry, S.T., Sirotkin, K., Starchenko, G., Suzek, T.O., Tatusov, R., Tatusova, T.A., Wagner, L. and Yaschenko, E: **Database resources of the National Center for Biotechnology Information.** Nucleic Acids Res. 2005. 33(Database issue):D39-45.

4. Gasteiger E., Jung E. and Bairoch A: **SWISS-PROT: Connecting biological knowledge via a protein database.** Curr. Issues Mol. Biol. 2001. 3:47-55.

5. Stenson, P.D., Ball, E.V., Mort, M., Phillips, A.D., Shiel, J.A., Thomas, N.S., Abeysinghe, S., Krawczak, M. and Cooper, D.N: **Human Gene Mutation Database (HGMD®): 2003 update.** *Human Mutation* 2003. 21(6): 577-581.

6. Wishart, D.S., Knox, C., Guo, A.C., Shrivastava, S., Hassanali, M. Stothard, P., Chang, Z. and Woolsey, J: **DrugBank: a comprehensive resource for in silico drug discovery and exploration.** Nucleic Acids Res. 2006 Jan 1;34(Database issue):D668-72.

7. Wishart, D.S., Tzur, D., Knox, C., Eisner, R., Guo, A.C., Young, N., Cheng, D., Jewell, K., Arndt, D., Sawhney, S., Fung, C., Nikolai, L., Lewis, M., Coutouly, M.-A., Forsythe, I., Tang, P., Shrivastava S., Jeroncic, K., Stothard, P., Amegbey, G., Block, D., Hau, David.D., Wagner, J., Miniaci, J., Clements, M., Gebremedhin, M., Guo, N., Zhang, Y., Duggan, G.E., MacInnis, G.D., Weljie, A.M., Dowlatabadi, R., Bamforth, F., Clive, D., Greiner, R., Li, L., Marrie, T., Sykes, B.D., Vogel, H.J., and Querengesser, L: **HMDB: the Human Metabolome Database.** Nucleic Acids Res. 2007. 35(suppl_1):D521-526.

8. Tanabe, L., Scherf, U., Smith, L.H., Lee, J.K., Hunter, L. and Weinstein, J.N: **MedMiner: an Internet text-mining tool for biomedical information, with application to gene expression profiling.** Biotechniques 1999. 27:1210-1217.

9. Hu, Y., Hines, L.M., Weng, H., Zuo, D., Rivera, M., Richardson, A. and LaBaer, J: **Analysis of genomic and proteomic data using advanced literature mining.** J Proteome Res. 2003. Jul-Aug;2(4):405-12.

10. Maier, H., Dohr, S., Grote, K., O'Keeffe, S., Werner, T., Hrabe de Angelis, M. and Schneider, R: **LitMiner and WikiGene: identifying problem-related key players of gene regulation using publication abstracts.** Nucleic Acids Res. 2005. Jul 1;33(Web Server issue):W779-82.

11. Hoffmann, R. and Valencia, A: **Implementing the iHOP concept for navigation of biomedical literature.** Bioinformatics 2005. 21 Suppl 2:ii252-ii258.

12. Plake, C., Schiemann, T., Pankalla, M., Hakenberg, J., and Leser, U: **Alibaba: PubMed as a graph.** Bioinformatics 2006. 22(19):2444-2445.

13. Rebholz-Schuhmann, D., Kirsch, H., Arregui, M., Gaudan, S., Riethoven, M. and Stoehr, P: **EBIMed – text crunching to gather facts for proteins from Medline.** Bioinformatics 2007. 23(2):e237-e244.

14. Schwartz, B., Melnikova, V.O., Tellez, C., Mourad-Zeidan, A., Blehm, K., Zhao, Y.J., McCarty, M., Adam, L., and Bar-Eli, M: **Loss of AP-2alpha results in deregulation of E-cadherin and MMP-9 and an increase in tumorigenicity of colon cancer cells in vivo.** Oncogene advance online publication 15 January 2007; doi: 10.1038/sj.onc.1210193.

15. Wain, H.M., Lush, M., Ducluzeau, F. and Povey, S: **Genew: the human gene nomenclature database.** Nucleic Acids Res 2002, 30:169-171.

16. Voorhees, E.M: **Overview of TREC 2005.** Proc. Text Retrieval Conf. (TREC), Gaithersburg, MD, November 2005, National Institute of Standards and Technology Special Publication 500-266.

17. Jones, K.S. and Rijsbergen, C.V: **Report on the need for and provision of an "ideal" information retrieval test collection.** British Library Research and Development Report 5266, Computer Laboratory, University of Cambridge, 1975.

18. Ruthven, I. and Lalmas, M: **A survey on the use of relevance feedback for information access systems.** The Knowledge Engineering Review, v.18 n.2, p.95-145, June 2003.

19. Rebhan, M., Chalifa-Caspi, V., Prilusky, J. and Lancet, D: **GeneCards: encyclopedia for genes, proteins and diseases.** Weizmann Institute of Science, Bioinformatics Unit and Genome Center. 1997.

20. MedGene Prostate Cancer Gene Evaluation: http://hipseq.med.harvard.edu/MedGene/publication/s_table3.html

21. Döhr, S., Klingenhoff, A., Maier, H., Hrabé de Angelis, M., Werner, T., and Schneider, R: **Linking disease-associated genes to regulatory networks via promoter organization.** Nucleic Acids Res. 2005. 33:864–872.

22. Hao, Y., Zhu, X., Huang, M., Li, M: **Discovering patterns to extract protein-protein interactions from the literature: PartII.** Bioinformatics. 2005. 21:3294-3300.

23. Entrez Programming Utilities: http://eutils.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html

24. Ding J., Berleant D., Nettleton D. and Wurtele E: **Mining MEDLINE: abstracts, sentences, or phrases?** Pac Symp Biocomput 2002. 326-37.

25. Hamosh, A., Scott, A.F., Amberger, J., Bocchini, C., Valle, D. and McKusick, V.A: **Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders.** Nucleic Acids Research 2002. 30:52-55.

26. Mishra, G., Suresh, M., Kumaran, K., Kannabiran, N., Suresh, S., Bala, P., Shivkumar, K., Anuradha, N., Reddy, R., Raghavan, T.M., Menon, S., Hanumanthu, G., Gupta, M., Upendran, S., Gupta, S., Mahesh, M., Jacob, B., Matthew, P., Chatterjee, P., Arun, K. S., Sharma, S., Chandrika, K. N., Deshpande, N., Palvankar, K., Raghavnath, R., Krishnakanth, K., Karathia, H., Rekha, B., Rashmi, N. S., Vishnupriya, G., Kumar, H. G. M., Nagini, M., Kumar, G. S. S., Jose, R., Deepthi, P., Mohan, S. S., Gandhi, T. K. B., Harsha, H. C., Deshpande, K. S., Sarker, M., Prasad, T. S. K. and Pandey, A: **Human Protein**

Reference Database - 2006 Update. Nucleic Acids Research. 2006. 34:D411-D414.

27. Becker, K.G., Barnes, K.C., Bright, T.J. and Wang, S.A: **The Genetic Association Database**. Nature Genetics 2004. 36: 431-432.

28. Thorisson, G.A., Smith, A.V., Krishnan, L. and Stein, L.D: **The International HapMap Project Web site.** Genome Res. 2005. Nov;15(11):1592-3

29. Riggins, G.J. and Strausberg, R.L: **Genome and genetic resources from the Cancer Genome Anatomy Project.** Hum Mol Genet. 2001. Apr;10(7)L663-7.

30. Humphreys, B.L., Lindberg, D.A., Schoolman, H.M. and Barnett, G.O: **The unified medical language system: An information research collaboration.** Journal of the American Medical Informatics Association. 1998. 5:1-13.

31. Brin, S. and Page, L: **The Anatomy of a Large-Scale Hypertextual Web Search Engine.** Computer Networks 30(1-7). 1998. 107-117.

32. Huang, M., Zhu, X., Hao, Y., Payan, D.G., Qu, K., Li, M: **Discovering patterns to extract protein-protein interactions from full texts.** Bioinformatics. 2004. 20:3604-3612.

33. Blaschke, C. and Valencia, A: **The potential use of SUISEKI as a protein interaction discovery tool.** Genome Inform ser Workshop Genome inform. 12:123-134

34. Knox, C., Shrivastava, S., Stothard, P., Eisner, R. and Wishart, D.S: **BioSpider: A Web Server for Automating Metabolome Annotations.** Pacific Symp. Biocomp. 2007, eds.

35. Rozen, S. and Skaletsky, H.J: **Primer3 on the WWW for general users and for biologist programmers.** Krawetz S, Misener S (eds) Bioinformatics Methods and Protocols: Methods in Molecular Biology. Humana Press, Totowa, NJ, pp 365-386.

36. IHOP Homo Sapiens dataset: http://www.ihop-net.org/UniPub/iHOP/info/gene_index/manual/1.html.

37. Witten, I. H. and Frank, E: Data Mining: **Practical machine learning tools and techniques.** 2nd Edition, Morgan Kaufmann, San Francisco, 2005.

38. The Merck Manual of Diagnosis and Therapy, Section 21. Special Subjects, Chapter 296. Normal Laboratory Values: http://www.merck.com/mmpe/index.html.

39. Google Scholar: http://scholar.google.com/.

40. Wikipedia: http://en.wikipedia.org/wiki/Main_Page.

41. Human Genome Project: http://www.ornl.gov/sci/techresources/Human_Genome/faq/snps.shtml

42. Gibbs, R.A. *et al.*: **The International HapMap Project.** Nature 2003. 426: 789-796.

43. Shastry, B.S: **Pharmacogenetics and the concept of individualized medicine.** Pharmacogenomics J. 2006. Jan-Feb;6(1):16-21.

44. McDonald, H. and Pereira, F: **Identifying gene and protein mentions in text using conditional random fields.** BMC Bioinformatics. 2005 May 24;6(Suppl I):S6.

45. Chen, H. and Sharp, B.M: **Content-rich biological network constructed by mining PubMed abstracts.** BMC Bioinformatics. 2004 Oct 8;5:147.

46. Radev, D.R., Hatzivassiloglou, V. and Mckeown, K.R: **A description of the CIDR system as used for TDT-2.** DARPA Broadcast News Workshop, Herndon, VA. 1999.

47. SIMP (SNP Information Mining Pipeline): http://bioinfo.bsd.uchicago.edu/SIMP.htm.